FIPS/T135

FAO Fisheries Technical Paper, No. 135

FIPS/T135
Statistical Methods

LIBRARY
NOV 14 1975
FISH & WILDLIFE SERVICE

APPLIED FISHERY STATISTICS

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS
ROME, 1974

Library
NOAA, National Marine Fisheries Service
SWFC Honolulu Laboratory F/SWC2
2570 Dole Street
Honolulu, HI 96822-2396

APPLIED FISHERY STATISTICS

by
G.P. Bazigos
FAO-Fishery Statistical Consultant
Fishery Statistics Unit
Policy and Planning Service

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS
Rome, 1974

Library
NOAA, National Marine Fisheries Servic
SWFC Honolulu Laboratory F/SWC2
2570 Dole Street
Honolulu, HI 96822-2396

PREPARATION OF THIS DOCUMENT

This report has been prepared in the Policy and Planning Service, Department of Fisheries, FAO, for training purposes in the field of fishery statistics (inland waters), in particular as a guide for the analysis of the results of large-scale fishery statistical surveys.

Distribution

FAO Department of Fisheries
FAO Statistics Division
FAO Regional Fishery Officers
FAO African Inland Waters Projects
Other FAO Fishery Projects
Specialized Institutes and
Individual Scientists

Bibliographic entry

Bazigos, G.P. (1974)
FAO Fish.Tech.Pap., (135):164 p.
Applied fishery statistics

Statistical methods - fisheries.
Statistics - theory and application.
Presentation of statistical data - tables and graphs. Statistical analysis - probabilities. Testing and tests. Analysis of variance. Experimental designs. Correlation and regression. Time series analysis. Fish yield indices. Simultaneous equations. Statistical tables. Selected bibliography.

CONTENTS

		Page
PREFACE		vii
ACKNOWLEDGEMENTS		viii
1.	TABULAR PRESENTATION OF STATISTICAL DATA	1
1.1	Classification of statistical data	1
1.2	Frequency distributions	2
1.3	Types of tables	3
1.4	The ingredients of a table	3
1.5	The tabulation system	5
2.	GRAPHIC PRESENTATION OF STATISTICAL DATA	5
2.1	Introduction	5
2.2	The system of coordinates	5
2.3	Fishery data shown by curves	6
2.4	Line diagrams for special purposes	7
2.4.1	Range charts	7
2.4.2	Component-part charts	8
2.5	Ratio or semi-logarithmic charts	8
2.6	Bar charts	10
2.7	Pie diagram	11
2.8	Charts of oblique projection	11
2.9	Statistical maps	12
2.9.1	Hatched maps	12
2.9.2	Dot maps	15
2.9.3	Other kinds of statistical maps	15
2.9.4	Isoline maps	15
2.10	Cumulative-frequency graphs	15
2.10.1	Ogive graph	15
2.10.2	Lorenz curve	18
2.11	Probability graphs	20
3.	MEASURES OF CENTRAL TENDENCY	22
3.1	The arithmetic mean (or average or mean)	22
3.2	The median	24
3.3	The quartiles, deciles	26
3.4	The mode (from grouped data)	28
3.5	Minor means	28
3.5.1	Geometric mean	28
4.	MEASURES OF SPREAD (Range, mean deviation, variance)	30
4.1	Tchebycheff's inequality	33
4.2	Gini's mean difference (d_1)	33
5.	MOMENTS	35
5.1	Moments	35
5.2	Higher moments	36
5.3	Measures of skewness	38
6.	PROBABILITIES IN A FINITE SAMPLE SPACE	39
6.1	Concepts and definitions	39

CONTENTS

		Page
6.2	Applications of probability	40
6.2.1	Illustrations	40
6.3	Combinatorial formulae	41
6.3.1	Illustrations of the use of combinatorial formulae	41
7.	THEORETICAL FREQUENCY DISTRIBUTIONS OF ONE VARIABLE	42
7.1	Random variables and their probability functions	42
7.2	Binomial distribution (discrete variable)	42
7.3	Poisson distribution (discrete variable)	44
7.4	Negative binomial distribution (discrete variable)	45
7.5	Normal distribution (continuous variable)	46
7.5.1	What is the normal distribution?	46
7.5.2	Properties of the normal curve	48
7.5.3	Parameters of the distribution	49
7.5.4	Cumulative curve	49
7.5.5	Sum and difference of normal variates	49
7.5.6	Standard normal distribution	49
8.	ESTIMATION AND TESTING HYPOTHESIS	51
8.1	Estimation	51
8.1.1	The statistical definition of a random sample	51
8.1.2	Estimators	51
8.1.3	Sampling distribution of the estimator t	52
8.1.4	Properties of the estimators	52
8.1.5	The estimator $\bar{x} = \frac{1}{n}\sum_i x_i$	52
8.2	Confidence interval for m (n>50)	53
8.3	Testing the hypothesis, Ho : $m = m_o$	54
8.4	Student's t variate	54
8.5	Confidence limits for a mean	55
8.6	Significance of the difference between two sample means $(\bar{x}_1-\bar{x}_2)$	56
8.7	Significance of the difference (m_1-m_2) when the population's variances are not known	57
8.8	Non-independent samples (t-student)	59
8.9	Distribution of proportion $p = \frac{x}{n}$ (large samples)	60
8.10	Distribution of the difference (p_1-p_2)	61
9.	THE χ^2 TEST	62
9.1	The χ^2 variate	62
9.2	The estimator of the variance σ^2	64
9.3	Significance of the difference between s^2 and σ^2	64
9.4	Confidence limits for σ^2	64
9.5	Testing goodness of fit	65
10.	DESCRIBING A FREQUENCY DISTRIBUTION BY A FITTED MATHEMATICAL MODEL	66
10.1	Fitting a normal curve	66
10.2	Fitting a binomial	67
10.3	Fitting a Poisson	68
11.	ANALYSIS OF VARIANCE	69
11.1	One criterion of classification	69
11.2	Two criteria of classification (one individual in each sub-class)	72

CONTENTS

	Page
11.3 Three criteria of classification (one individual in each sub-class)	75
11.4 Analysis of covariance	81
11.4.1 One-way classification	81
11.4.2 Two-way classification	85
12. EXPERIMENTAL DESIGNS	85
12.1 Introduction	85
12.2 Randomization, replication and selection of treatments	85
12.3 The latin square	86
12.4 Factorial experiments	88
13. CORRELATION AND REGRESSION	91
13.1 Introduction	91
13.2 Correlation (bivariate distributions)	91
13.2.1 Linear correlation	92
13.2.2 Rank correlation (non-parametric methods)	96
13.2.3 Correlation of grouped data	96
13.3 Linear regression	97
13.3.1 The regression line of y on x	98
13.3.2 The mean square error	100
13.3.3 The standard error of regression coefficient	101
13.4 Further linear correlation and regression theory	101
13.4.1 The correlation coefficient and explained variation	101
13.4.2 Fiducial limits of y_c for given x_o	104
13.4.3 Correlation ratio	105
13.4.4 The linear regression of x on y	106
13.5 Linear multiple regression	107
13.5.1 Introduction	107
13.5.2 Partial correlation and multivariate regression	107
13.5.3 The type of equation (linear multiple regression)	107
13.5.4 The system of normal equations	107
13.5.5 Gauss multipliers	109
13.5.6 Mean square error	110
13.5.7 Analysis of total variation into components	110
13.5.8 Estimated confidence limits for the population partial regression coefficients	111
13.5.9 The significance of the individual variables	111
13.5.10 The multiple correlation coefficient	112
13.5.11 The reliability of the multiple correlation coefficient	113
13.5.12 Partial correlation coefficients	113
13.5.13 The reliability of partial correlation coefficients	114
13.6 Polynomial regression	115
13.7 Regression of equations using transformations	118
13.7.1 The log y, log x relationship	119
13.7.2 The log y, x relationship	120
13.7.3 The $\frac{x}{y}$, x relationship	121
13.7.4 The $\frac{1}{y}$, x relationship	122
13.7.5 Use of transformations in multiple curvilinear regression	122
14. METHODOLOGY OF FISHERY TIME SERIES	126
14.1 Introduction	126

CONTENTS

	Page
14.2 Introducing time as an explanatory variable	126
14.3 Basic trends	126
14.3.1 Basic trend 1 : the straight line	126
14.3.2 Basic trend 2 : second-degree curve	128
14.3.3 Basic trend 3 : compound-interest curve	130
14.3.4 Basic trend 4 : modified exponential curve	132
14.3.5 Basic trend 5 : Gompertz curve	133
14.4 Correlation of fishery time series	136
14.5 Multiple estimating equations	138
15. FISH YIELD INDICES	138
15.1 Introduction	138
15.2 The construction of the index	139
15.2.1 The population of species	139
15.2.2 Selection of the base	140
15.2.3 The data for the indices	140
15.2.4 Yield indices defined	140
15.2.5 The statistical formulae of estimation	141
APPENDIX I - MATHEMATICS OF SIMULTANEOUS EQUATIONS	147
1. SIMULTANEOUS LINEAR EQUATIONS IN TWO UNKNOWNS	147
2. SIMULTANEOUS LINEAR EQUATIONS IN THREE UNKNOWNS	147
3. DETERMINANTS	148
3.1 Application of determinants to simultaneous equations	150
APPENDIX II - STATISTICAL TABLES	153
REFERENCES	162
LIST OF TECHNICAL REPORTS	163

PREFACE

In the field of applied fishery researches at inland waters in Africa where FAO Projects are operating the correct interpretation of data is of supreme importance, since, unless correct conclusions can be drawn from them, the acquisition of numerical results has little or no value. In this connection it has been recognized that statistical methods provide an indispensable tool for solving many of the problems which are encountered in applied fishery studies. The purpose of this manual is to help the fishery statisticians working in the field and other interested researchers to acquire knowledge of the subject and to emphasize the economies to be achieved from the intelligent application of the proper statistical methods. One, however, should remember that we do need reliable statistical data to make the right type of analysis.

ACKNOWLEDGEMENTS

The author of a manual is grateful to almost anyone who has touched the field and I take pleasure in acknowledging my debt to all those persons, named or unnamed, whose results I have drawn upon during the writing of this manual which is the product of five years experience at large African lakes where FAO Projects are being undertaken.

My first acknowledgement must go to Mr. C.H. Clay, FAO Co-ordinator of African Lakes Projects, for the unstinting assistance given in connection with my assignments. The manual could not have been written without the opportunity, made possible by the Managers of FAO Projects, my African counterparts in the field and other field staff, to use statistical methods for analyzing the results of large-scale surveys, for which I shall ever remain grateful.

I would also like to record my thanks to Messrs. S.R. Coppola and H. Thiessen for their assistance required for the preparation of the manual, and last but not least my thanks are due to Miss Sheila Campbell for the patience with which she typed the drafts and master copy.

1. TABULAR PRESENTATION OF STATISTICAL DATA

In this chapter we discuss methods of tabulation of fishery statistical data obtained through a frame survey or a sample survey[1/].

1.1 Classification of statistical data

Generally speaking, there are four important bases for the classification of statistical data: a) qualitative, b) quantitative, c) geographical, and d) chronological, each of which is briefly examined in turn.

Qualitative: The classification is one of kind rather than of amount. Fishermen may be classified concerning employment status, as boat-owners, with gear only and assistants. Fishing boats may be classified as plank boats, canoes and catamarans, etc.

Quantitative: When items vary in respect to some measurable characteristic, a quantitative classification is appropriate. Fishing economic units may be classified according to the number of active fishermen. Fishing sites may be classified according to the number of local fishing boats, etc. Most quantitative distributions are frequency distributions (see section 1.2).

Geographical: The geographical distribution is essentially a type of qualitative distribution. Here the control characteristic for classification of the survey units is the geographical location of the units. When fishing boats are shown on a limnological zone basis, we have data which are classified geographically. Sometimes a geographical distribution may be put into the form of a frequency distribution:

Area distribution of fishing boats at lake L.

Place	Fishing boats		Remarks
	Number	(%)	
Total	8000	100.00	
LZ-1	2100	26.25	(LZ : Limnological zone)
LZ-2	1800	22.50	
LZ-3	1300	16.25	
LZ-4	2500	31.25	
LZ-5	300	3.75	

Chronological: Chronological data or time series show figures concerning a particular phenomenon at various specified times. For example, fish catch data (lake L) may be shown monthly, quarterly or annually. The essential arrangement of the figures in a time sequence is inherent in the nature of the data under consideration.

Occasionally a time series may be converted into a frequency distribution:

1/ For the collection of statistical data in the field of fishery statistics see, G.P. Bazigos: The Design of Fisheries Statistical Surveys - Inland Waters, FAO, Fish.Tech.Pap. FIPS/T133. May, 1974.

Age distribution of fishing boats at lake L

Age of fishing boats (months)	Fishing boats		Remarks
	Number	(%)	
Total	8000	100.00	
1-6	3200	40.00	
7-12	1600	20.00	
13-18	1200	15.00	
19-24	800	10.00	
25-30	800	10.00	
31-36	400	5.00	

1.2 Frequency distributions

A large set of numbers e.g. values of a continuous variate[1/], is presented in the form of a frequency distribution by dividing the range of numbers into intervals (usually equal) and setting down the number of individual values in each interval; this process is called grouping and the number in each interval is called the frequency of the interval. The relative interval frequency is this frequency divided by the total frequency.

Example: The table below shows the grouping of 150 measurements of the continuous variate x; length of fish, species A (class interval 1cm).

x		
10-11	||	2
11-12	𝍸 𝍸 |||	13
12-13	𝍸 𝍸 𝍸 𝍸 |	21
13-14	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |	36
14-15	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ||||	34
15-16	𝍸 𝍸 𝍸 𝍸 𝍸	25
16-17	𝍸 𝍸 ||	12
17-18	𝍸 ||	7

The results can be presented as follows:

Centre of interval x_i	10.5	11.5	12.5	13.5	14.5	15.5	16.5	17.5	Total
Frequency f_i	2	13	21	36	34	25	12	7	$\sum_i f_i = 150$
f_i (%)	1.33	8.67	14.00	24.00	22.67	16.67	8.00	4.66	100.00

In the case of a discrete variate the frequency distribution can be presented as follows:

1/ Data of a discrete variate are such as number of active fishermen per fishing economic unit, number of gillnets owned per unit, etc. Data of a continuous variate are such as fishing time, fish catch (weight), length of fish, etc.

x_i	x_1	x_2	x_3	....	x_i	....	x_k	Total
f_i	f_1	f_2	f_3	....	f_i	....	f_k	$\sum_i f_i$
f_i (%)	f_i (%)	f_2 (%)	f_3 (%)	....	f_i (%)	....	f_k (%)	100.00

1.3 Types of tables[1/]

From the point of view of presentation of statistical data, there are two types of tables. In the first place there are the detailed tables or reference tables, giving detailed information of the survey variates and arranged for ready reference. In the second place there are the summary or text tables, which are usually relatively small in size and which are designed to set forth one finding or a few closely related findings as effectively as possible. The tabulated data of a summary table are the totals or sub-totals contained in one or more reference tables, although upon occasion a summary table may be based, in whole or in part, upon one or more other summary tables.

The arrangement of the items of information in a table is a function of the nature of the data. However, the needs of the consumers of the statistical data should be taken into account. A good table presents the data in such a way that it simplifies a further statistical analysis of the tabulated data. General statistics points out a number of different methods for tabulation of statistical data i.e.

Alphabetical: In detailed tables the survey characteristics are usually arranged in alphabetical order because it enables individual items of information to be located with ease.

Geographical: Any geographic arrangement of statistical data should be supported with the proper mapping material in order to simplify the task of the consumers of the data.

Magnitude: In certain cases it is desirable to arrange items of information in summary tables according to their size, usually with the largest item first, but sometimes with the order reversed.

Historical: Data of time series should be arranged historically. When years are listed, either the most recent or the earliest date may be shown first. The months, however, are customarily listed with January first.

Customary: Qualitative data are arranged by taking into account the classes of the survey characteristics as well as the existing interrelationship between them.[2/]

1.4 The ingredients of a table

The main ingredients of a table are: a) The heading of the table, b) The body of the table, and c) Notes (prefatory notes, footnotes, and source notes).

Heading: The heading of a table includes the serial number and the title of the table stating in order: what, where, how classified, and when.

1/ See also, the processing of the results of sample surveys in the manual given in footnote 1/, page 1.

2/ Various types of statistical tables are given in, G.P. Bazigos: Frame Survey at Kainji Lake, St.S./1, FI/SF/NIR/24, January 1971.

Body: Includes the items of information presented by the table. The individual items of information are placed in columns (rows) with clear short headings.

Notes: A prefatory note is placed just below the heading of a table and in smaller or less prominent type and provides an explanation concerning the entire table or a substantial part of it. Explanations concerning individual figures or a column or row of figures should be given in footnotes. Source notes may appear below the title or below the footnotes.

Examples of tables are given below:

Table ... Estimated average weight of fish caught per ha (kg/ha) and ICC's*/ by stratum and for the lake as a whole, fishing year July 1970-June 1971

Space	Estimated average weight of fish, kg/ha		Estimated ICC's
	Totality of species	Citharinus spp.	
Total	223.7	176.3	78.8
Stratum I	185.1	165.4	89.4
Stratum II	201.0	171.9	85.5
Stratum III	479.1	233.2	48.7

*/ ICC : Ichthyomass Concentration Coefficient

Source : Catch Assessment Survey, Lake Kainji, Nigeria

Table ... The transport of fish by fishermen and fishmongers

Method of transport	Number of canoes responsible for the commercial production							
	Fishermen				Fishmongers			
	First movement		Second movement		First movement		Second movement	
		(%)		(%)		(%)		(%)
Total	2022	100.0	58	100.0	1393	100.00	957	100.0
1. Headload	714	35.3	52	89.6	364	26.1	105	11.0
2. Canoe	1177	58.2	-	-	682	49.0	-	-
3. Bicycle	29	1.4	-	-	17	1.2	40	4.2
4. Lorry	102	5.1	6	10.4	330	23.7	812	84.8

Source : Frame Survey (September 1970), Lake Kainji, Nigeria

Table ... Estimated values of control time series at Lake Nasser, 1972-76[1]/

Year	Expected surface area of the lake (ha)	Expected commercial landings (metric tons)		Expected values of $\hat{\beta}_1$ ($\hat{\beta}_1$ = kg/ha/yr)	
		Minimum	Maximum	Minimum	Maximum
1972	342500	8860	9600	25.87	28.03
1973	355000	11684	12893	32.91	36.32
1974	415000	15328	15807	36.93	38.09
1975	445000	17498	18028	39.32	40.51
1976	475000	19710	20289	41.49	42.71

1/ G.P. Bazigos, The Yield Pattern at Lake Nasser, St.S./1, UNDP/SF/EGY/66/558, September 1972.

1.5 The tabulation system

The tabulation system gives the complete set of blank tables to be used for the presentation of the results of a survey. In the system, by using the proper codification system, the type of the tables, the group number and the serial number of each table is provided. Also, the population of the tabulated characteristics and their interrelationship in the system are portrayed. It should be noted that, in case of mechanical processing of the data, the tabulation system of the survey must be discussed in advance with the programmer. The preparation of the tabulation system of the survey is the responsibility of the Fishery Statistician. The programmer examines the operational requirements of the proposed system and comments on it. A tabulation system is considered to be good if it demonstrates simplicity, flexibility and low processing cost and at the same time presents in a systematic way all the information collected through a statistical survey.

2. GRAPHIC PRESENTATION OF STATISTICAL DATA

2.1 Introduction

This chapter is devoted to a discussion of the portrayal of fishery statistical data by graphic devices. The outstanding effectiveness of charts is that they make possible the presentation of quantitative data in a simple, clear, and effective manner and facilitate comparison of values, trends, and relationships. One should remember that, although graphic techniques are a powerful and effective medium for presenting statistical data, they are not under all circumstances and for all purposes complete substitutes for tabular and other forms of presentation.

2.2 The system of coordinates

The system of coordinates which has a good use in graphic presentation consists of a pair of intersecting lines, called axes, which are drawn at right angles. The horizontal line is referred to as the X-axis, or axis of abscissas, and the vertical, the Y-axis, or axis of ordinates. These axes divide the region of the plane into four compartments, called quadrants, which are numbered counterclockwise, beginning with the upper-right quadrant.

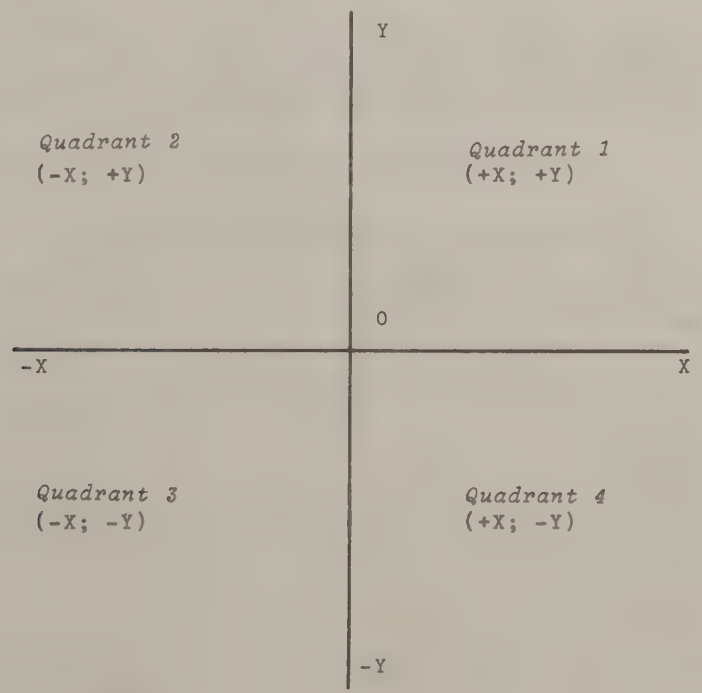


Figure 2.2.1

2.3 Fishery data shown by curves

In the field of fishery statistics curves are frequently used for picturing a) time series, i.e. either period data over time such as yearly fish catch, quarterly fish catch, etc, or point data over time such as active fishermen at different points of time (survey date) etc, and b) frequency distributions.

Whenever chronological data are depicted by means of a curve, the years, quarters, or other chronological units are shown on the horizontal axis; the other series which varies with time is placed on the vertical axis. In the figure below (Figure 2.3.1) the chronogram indicates the relative changes in the total number of fishing boats on a lake over time (1969-73).

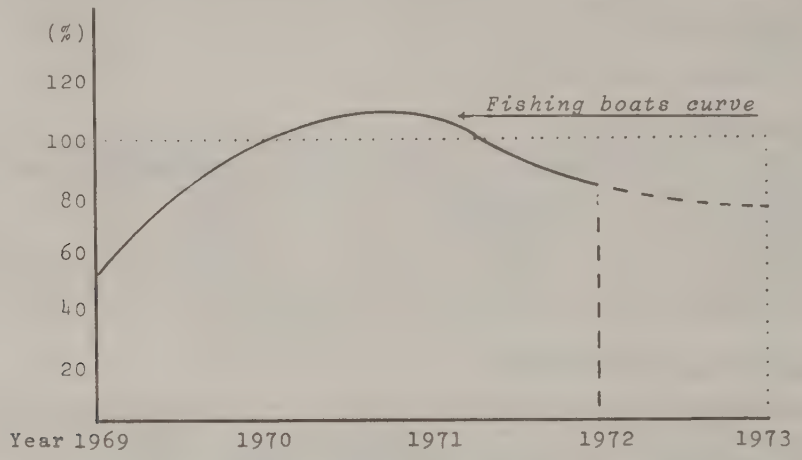


Figure 2.3.1 Relative changes in the total number of fishing boats on a lake over time (results of a Current Frame Survey), 1970 = 100

Continuous frequency distributions are portrayed by histograms. The table below gives the frequency distribution of the fishing camps at lake A according to the average total daily fish catch. Prepare the respective histograms by using the absolute frequencies (Figure 2.3.2a) and the relative frequencies (Figure 2.3.2b).

Table 2.3.1 Frequency distribution of the fishing camps at lake A according to the average total daily fish catch

Average total daily fish catch kg (x)	No. of fishing camps f_i	f_i (%)	Remarks
0-20	10	5.00	
20-40	10	5.00	
40-60	30	15.00	
60-80	80	40.00	
80-100	30	15.00	
100-120	20	10.00	
120-140	10	5.00	
140-160	5	2.50	
160-180	5	2.50	
	200	100.00	

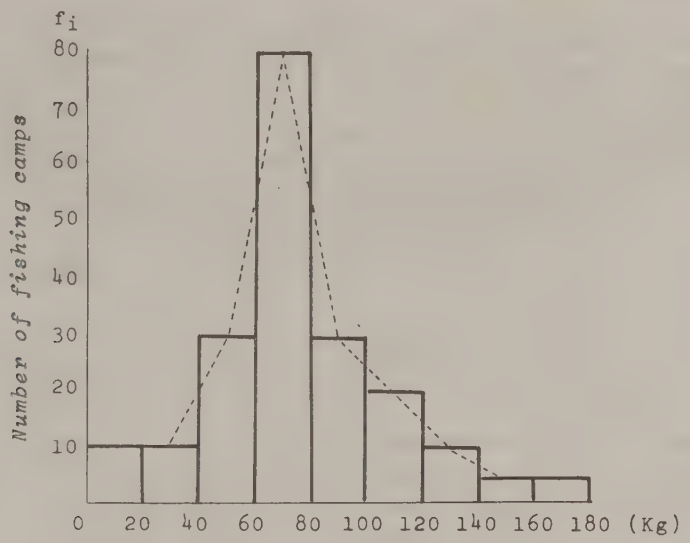


Figure 2.3.2a Absolute frequencies

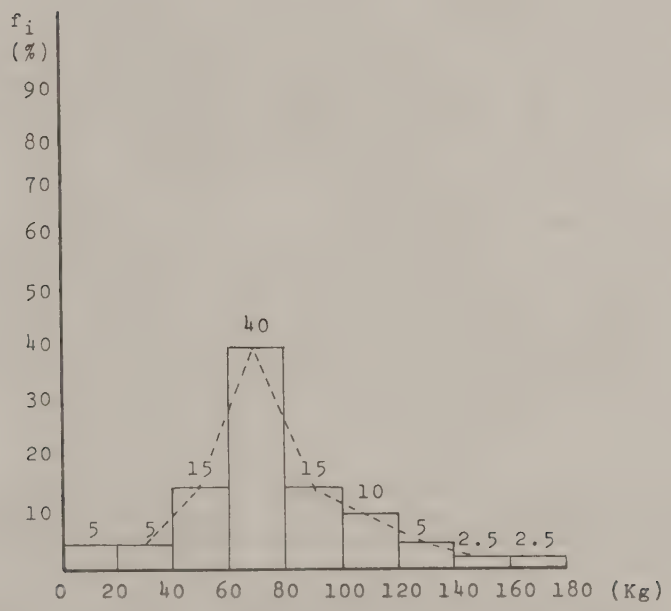


Figure 2.3.2b Relative frequencies

2.4 Line diagrams for special purposes

2.4.1 Range charts

Range charts are used in the field of fishery statistics for showing the range of variations between parallel magnitudes over time. In the following figure (Figure 2.4.1.1) the range chart indicates the range of variations between total fish catch and commercial fish catch[1/] over time at lake A.

1/ Commercial fish catch is a part of the total fish catch, which enters the distribution channel within a period of time.

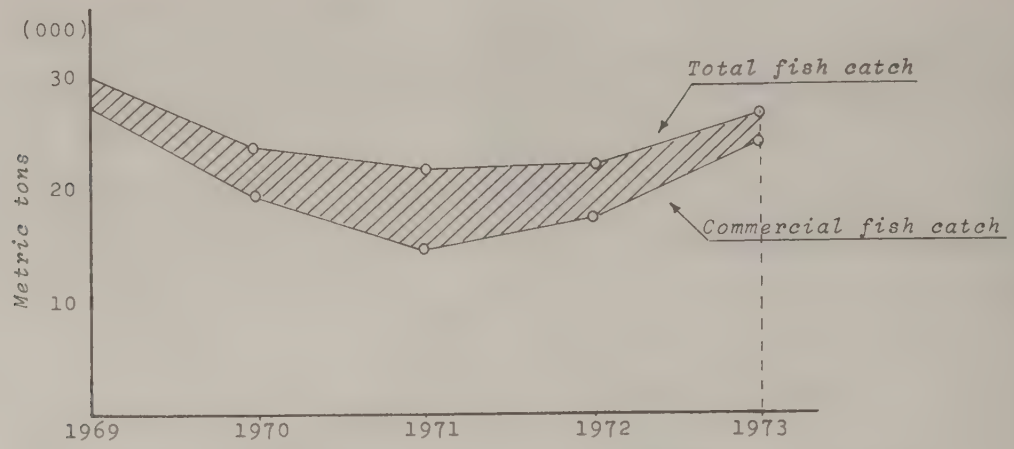


Figure 2.4.1.1 Variations between total fish catch and commercial fish catch

2.4.2 Component-part charts

Component-part charts are widely used in fishery statistics, among other things, to portray species composition of the fish catch over time. It is possible to observe, from this type of chart, whether or not a given species (group of species) is increasing or decreasing over time, and whether or not the total fish catch is increasing or decreasing. The figure below (Figure 2.4.2.1) shows the relative composition of fish catch over time at Lake Volta, 1969, 1970.

2.5 Ratio or semi-logarithmic charts

Semi-logarithmic charts are used in the field of fishery statistics to represent the ratio of change that occured between two points of time or between successive points in time. In this kind of chart[1/] the vertical axis is ruled logarithmically, and the horizontal axis, arithmetically. It should be remembered that an arithmetic chart portrays the absolute change between two point values, i.e. $d = x_2 - x_1$. A semi-logarithmic chart gives the logarithm of the ratio,

$$\log x_2 - \log x_1 = \log\left(\frac{x_2}{x_1}\right)$$

The essential differences between arithmetic and semi-logarithmic charts can be illustrated through the following example. The average catches per fishing operation in two limnological zones (LZ-1, LZ-2) of a lake during the dry and rainy periods are given in the table below,

Limnological zone	Fishing period		Absolute diff. between fishing periods	Relative diff. between fishing periods	Remarks
	Dry (kg)	Rainy (kg)			
LZ-1	10	11	1kg	10%	
LZ-2	90	99	9kg	10%	

1/ Semi-logarithmic charts and logarithmic charts of various cycles are commercially available.

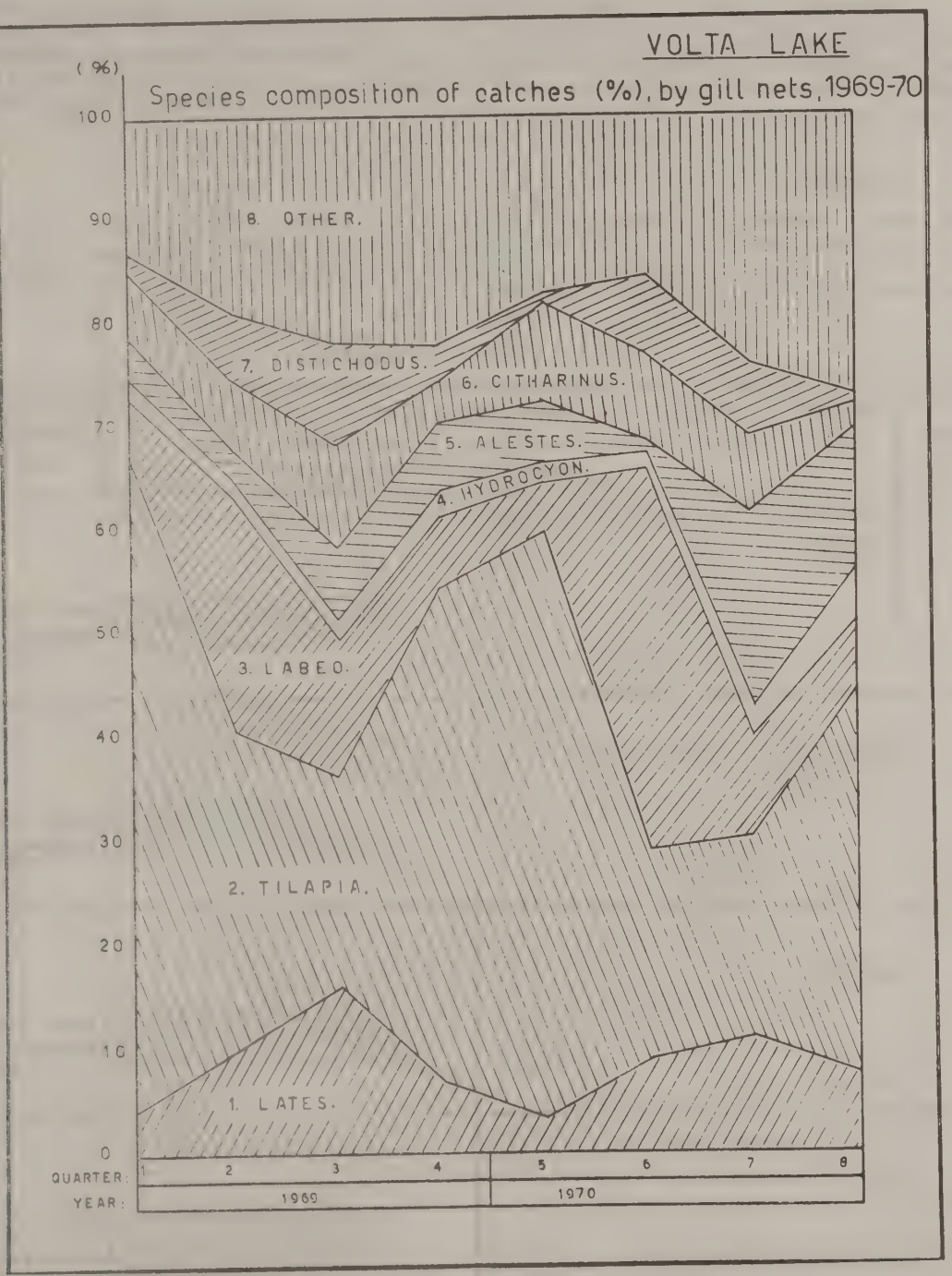


Figure 2.4.2.1 Relative composition of fish catch over time at Lake Volta, 1969, 1970

On an arithmetic chart the amount of increment or change for the first pair (LZ-1) would be one unit and for the second pair (LZ-2) would be nine units. This in turn means that the upward movement of the curve for the second pair on an arithmetic grid would be nine times as great as the upward movement of the curve for the first pair of figures. By contrast, the curves for these data when plotted on a semi-logarithmic chart would have identical slopes, because the relative or percentage changes are the same.

2.6 Bar charts

Bar charts (vertical or horizontal) are used in fishery statistics either to portray chronological data or for comparison of data which is classified qualitatively or quantitatively. The following figure (Figure 2.6.1) shows the relative composition of fishing labour force at Lake Kainji (Frame Survey 1).

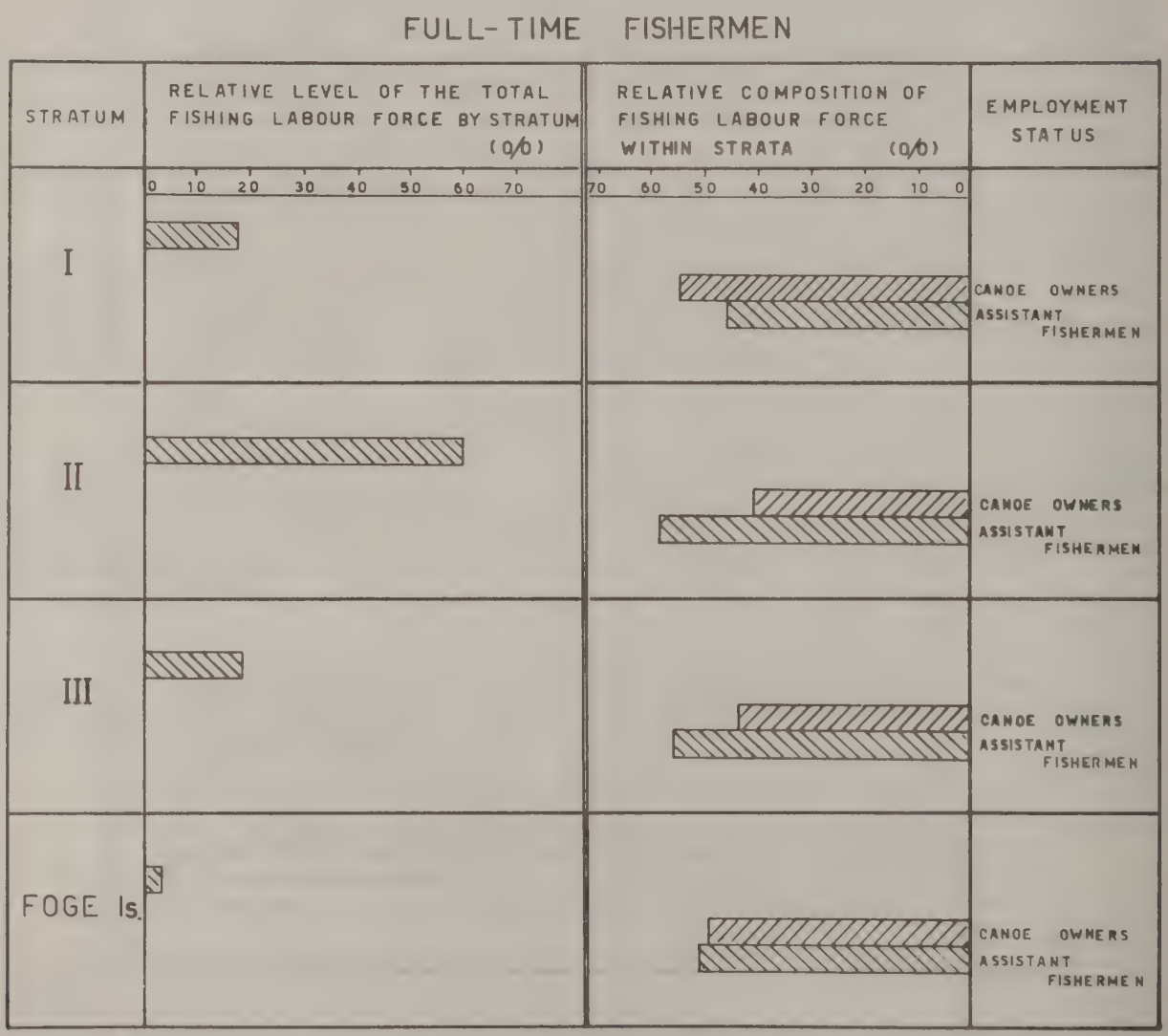


Figure 2.6.1 Relative composition of fishing labour force at Lake Kainji

2.7 Pie diagram

Pie charts are used to portray component relations. The various sectors of a circle represent component parts of an aggregate or total. For constructing a pie chart the following steps should be taken:

1. The various component values must be transposed into corresponding degrees of the circle. For example, the yearly fish catch composed of three species, A:30 percent, B:50 percent, C:20 percent. Since 1 percent is equal to 3.6 degrees (360/100 = 3.6°) then the corresponding values for the three species are 30x3.6° = 108°, 50x3.6° = 180°, 20x3.6° = 72°.
2. The second step is to plot a circle of appropriate size.
3. Points on the circle representing the size of a sector are measured with a protractor.

The Pie diagram below (Figure 2.7.1) portrays the relative distribution of the yearly total fish catch on a stratum basis (fishing year 1970 / 71) at Lake Kainji.

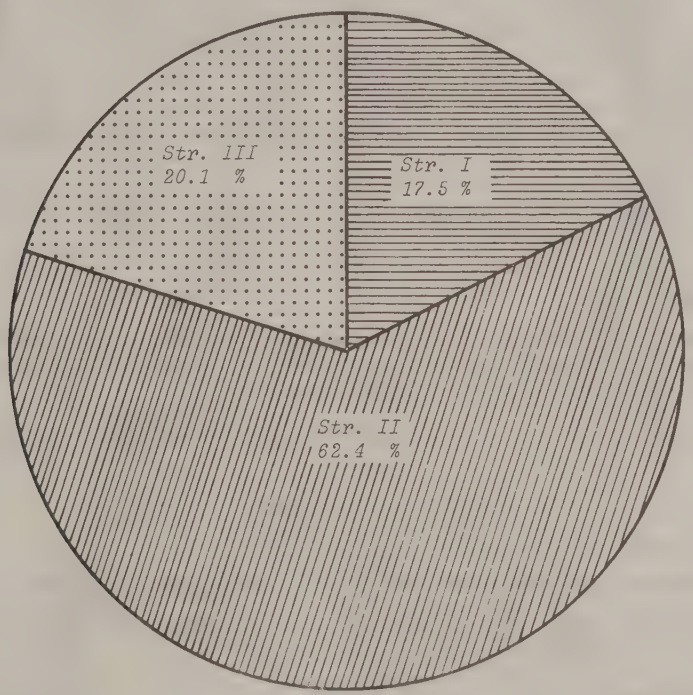


Figure 2.7.1 The relative distribution of the yearly total fish catch on a stratum basis, fishing year 1970/71

2.8 Charts of oblique projection

Charts of oblique projection are mainly used in fishery statistics to portray the level of dispersion of the survey units. The chart below indicates the dispersion pattern of pelagic fish in Lake Tanganyika. The values of the bars of L_1 (LZ-1) can be read simply on the y-axis, whereas, the values of the bars of L_2 (LZ-2) and L_3 (LZ-3) can be read on the y-axis only by taking into account the correspondingly higher drawn axes of L_2 and L_3. The numbers in the chart are the values of the estimated specific indices expressing relative changes of the survey variate between the established waterroads, (LZ : limnological zone).

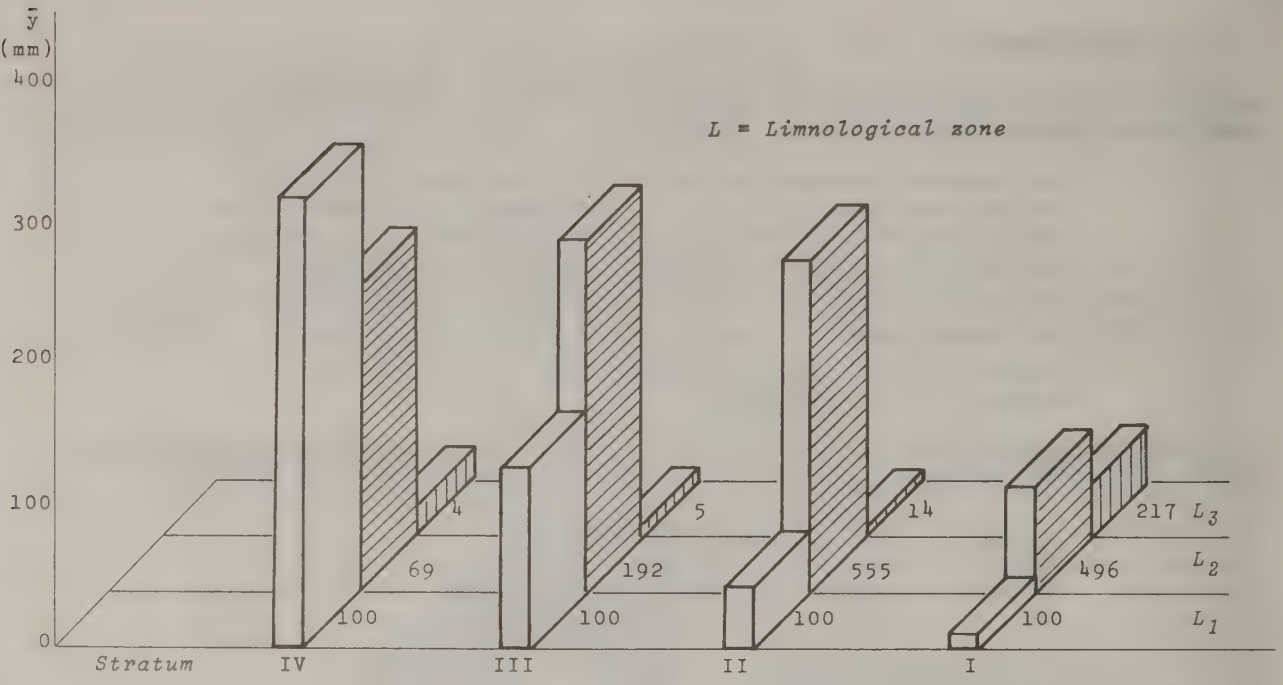


Figure 2.8.1 Dispersion pattern of pelagic fish in Lake Tanganyika, November 1973

2.9 Statistical maps

The most effective technique for portraying spatial relationships is the map. Specifically, statistical maps are graphic devices which show qualitative as well as quantitative information on a geographical basis.

In constructing a statistical map consideration should be given to the following points:

1. The choice of the scale of the map.
2. The accuracy of the map.
3. The amount of information to be placed on the map.

In this section we shall consider types of statistical maps which are in use in the field of fishery statistics.

2.9.1 Hatched maps

Generally speaking, hatched maps or shaded maps undertake to show for each geographical area under consideration the magnitude of the survey characteristic. The variations in magnitude are presented graphically by progressive differences in hatching or shading. Specifically, in the field of fishery statistics and biology, hatched maps are mainly used to portray:

1. The relative abundance of fish in a lake.
2. Variations in the area distribution of fishing effort, etc.

In Figure 2.9.1.1 the hatched map portrays the relative abundance of pelagic fish at Lake Tanganyika (results of an echo-survey). In Figure 2.9.1.2 the hatched map portrays variations in the area distribution of fishing effort at Lake Kainji (Nigeria) (results of a Frame Survey).

Figure 2.9.1.1 The relative abundance of fish in Lake Tanganyika[1/]

1/ Johannesson, K.A., Preliminary quantitative estimates of pelagic fish stocks in Lake Tanganyika by use of echo integration methods, EIFAC/74/I Symp.54.

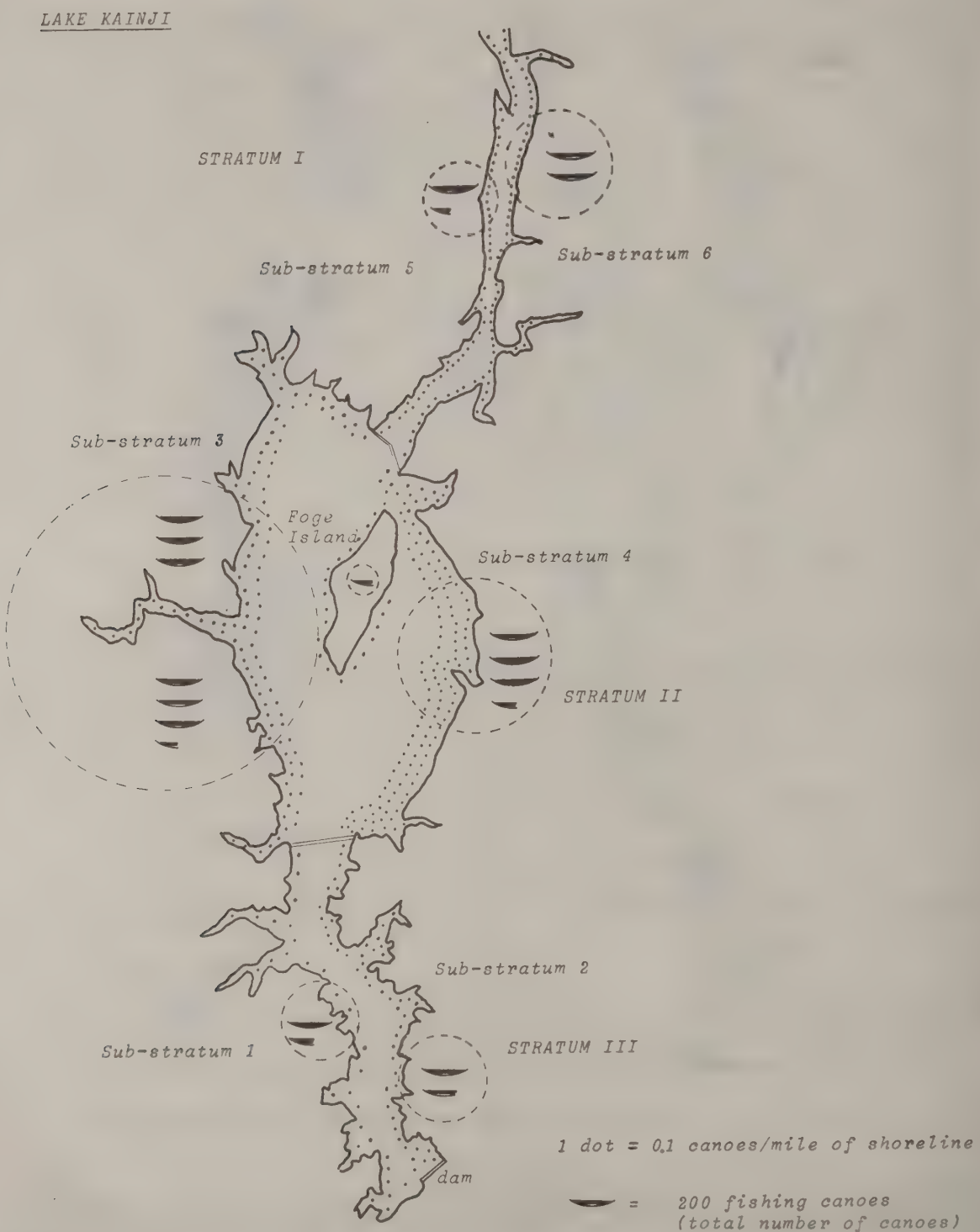

LAKE KAINJI
STRATUM I
Sub-stratum 5
Sub-stratum 6
Sub-stratum 3
Foge Island
Sub-stratum 4
STRATUM II
Sub-stratum 2
Sub-stratum 1
STRATUM III
dam
1 dot = 0.1 canoes/mile of shoreline
= 200 fishing canoes (total number of canoes)

Figure 2.9.1.2

2.9.2 Dot maps

Dot maps are mainly used to portray geographical distribution of occurrences i.e. emphasis is placed on absolute number. Specifically, in the field of biology and fishery statistics, dot maps are mainly used to portray:

1. Absolute abundance of fish on a regional basis e.g. kg/ha. Here, one dot might correspond to one kg of fish or 10 kg of fish, etc.
2. Regional fish catch in a lake, e.g. m.tons/year. Here, one dot corresponds to 1 ton of fish or 1000 tons of fish, etc. (see Figure 2.9.2.1).

2.9.3 Other kinds of statistical maps[1/]

In the field of fishery statistics "pin maps" are used to portray the area distribution of FEU's[2/] according to their fishing method used. For this purpose different colours of pins should be used, e.g.:

FEU's using gillnet fishing: black pins.

FEU's using Lusenga fishing: red pins.

FEU's using beach seine fishing: blue pins, etc.

It is obvious that by using such maps it is possible to observe both the general and specific (by fishing method) concentration of fishing effort in space.

Maps with various symbols are used to portray the centres providing capital goods to a fishing industry at a lake, i.e. fishing boats and fishing gear.

Maps with "directing lines" are used to portray the peripheral movement of the FEU's and the migration pattern of the fishermen.

2.9.4 Isoline maps

Some of the more familiar examples of isoline maps in the field of applied fishery researches include those pertaining to a series of points of equal temperature in a lake, which are called "isotherms" or equal elevation "isohypses" or, more commonly, "contour lines", etc. The basic data used in the preparation of isoline maps may be samples of absolute measurement for particular points or ratios for certain areas. In both cases points having identical values are joined together by continuous flowing lines.

2.10 Cumulative-frequency graphs

2.10.1 Ogive graph

For some purposes the cumulative-frequency curve, or ogive, is more valuable than the simple frequency graph. As we have discussed, in order to show the genesis of a simple frequency curve, the data of a frequency distribution are presented by a series of rectangles or bars in the "column diagram" and this diagram is transformed into a curve by connecting the midpoint on the top of each rectangle with the midpoint of each adjacent rectangle.

For the construction of an ogive graph the following steps are taken:

1. We calculate the cumulative frequencies of the frequency distribution. For the accumulation process we can move either from the lesser to the greater, referred to as a "less than" type of distribution, or from the greater to the lesser, known as a "more than" type of distribution.

1/ See, Frame Survey at Kainji Lake, St.S./1, FI/SF/NIR/24, January 1971, G.P. Bazigos.

2/ FEU's : Fishing Economic Units.

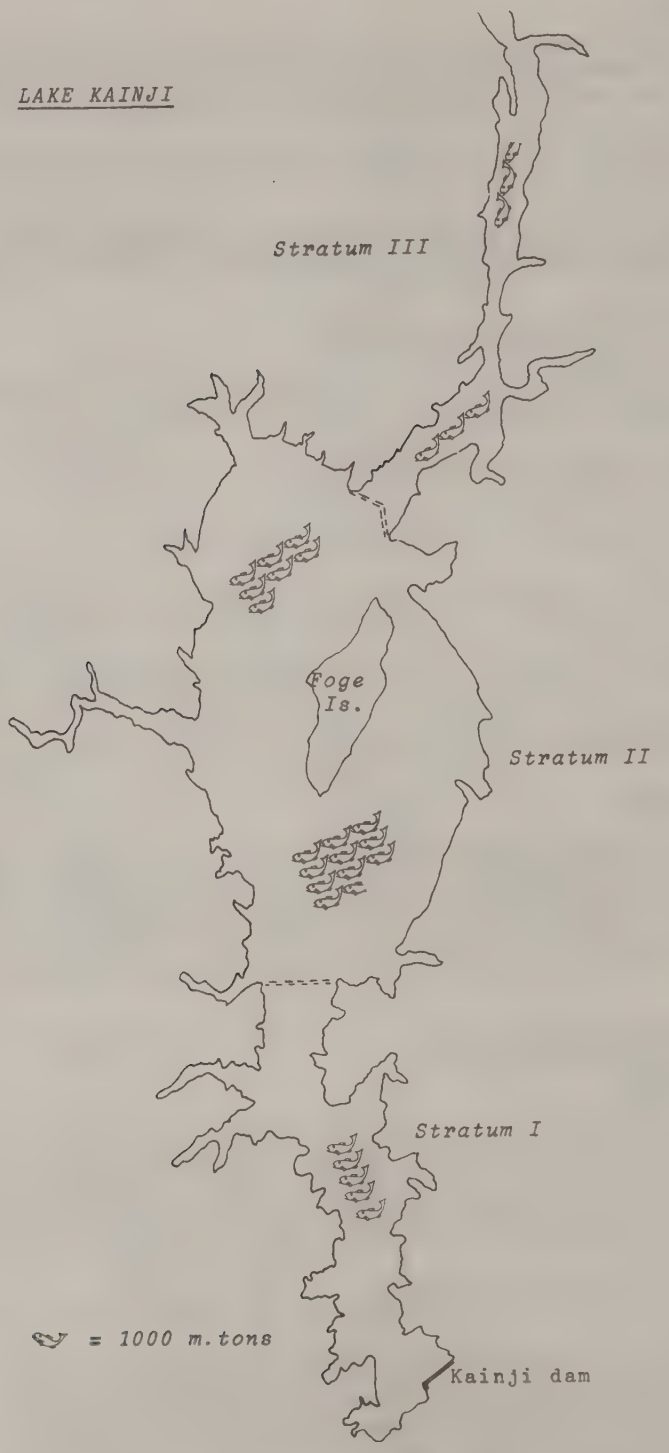


Figure 2.9.2.1 Estimated total fish catch by stratum at Lake Kainji, 1970-71, CAS

2. In the horizontal axis (X) we represent the class intervals and in the vertical axis (Y) the cumulative frequencies.
3. We plot the cumulative frequencies either in the upper or lower end of the respective class interval, depending on whether the accumulation is of the "less than" or "more than" type.
4. Ogive graphs can be produced for absolute cumulative frequencies and relative cumulative frequencies (percentages).
5. Both the numerical and percentage cumulative-frequency curves are of the characteristically elongated S shape.

Ogive graphs are of great value in the field of applied fishery statistics for describing the organic structure of the survey populations. In the table below, information is provided on the distribution of the small size fishing sites at lake A according to the average total output per day (kg). The produced ogive graphs describing the catch efficiency of the fishing sites of the industry are given in Figure 2.10.1.1a, b.

Table 2.10.1.1 Catch efficiency distribution of the small size fishing sites at lake A

Kg/day	No.of fishing sites	F_i	F'_i	F_i (%)	F'_i (%)	Remarks
0-5	65	65	765	8.50	100.00	
5-10	190	255	700	33.33	91.50	
10-15	235	490	510	64.05	66.67	
15-20	165	655	275	85.62	35.95	
20-25	90	745	110	97.39	14.38	
25-30	15	760	20	99.35	2.61	
30-35	5	765	5	100.00	0.65	
	765					

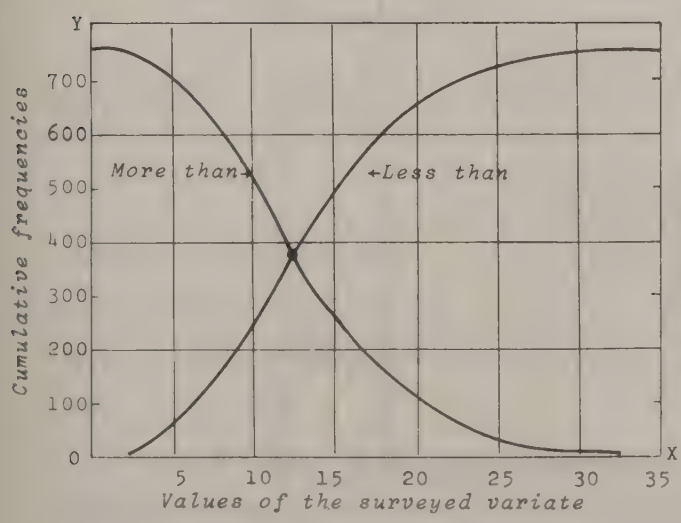


Figure 2.10.1.1a

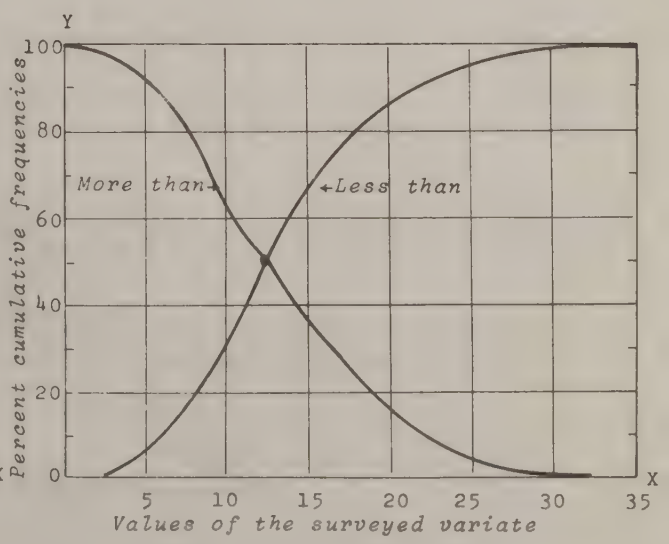


Figure 2.10.1.1b

It should be noted that in comparing ogives emphasis is placed on the number of proportion of cases above or below a given point. It should also be remembered that the maximum frequency or concentration of cases on an ogive is indicated by the steepest portion of the curve.

2.10.2 Lorenz curve

The Lorenz curve is a special type of cumulative-frequency graph and it has a good deal of application in the field of economic statistics. The table below (Table 2.10.2.1) provides the basic steps for the construction of a Lorenz curve.

Table 2.10.2.1 Size distribution of the fishing camps at lake A

No.of fishing boats x_i (1)	No.of fishing camps f_i (2)	Total No. of fishing boats $f_i x_i$ (3)	$F_i = \sum_i f_i$ (4)	$\phi_i = \sum_i f_i x_i$ (5)	$F'_i = \sum_i f_i / N_1$ (%) (6)	$\phi'_i = \sum_i f_i x_i / N_2$ (%) (7)
1	21	21	21	21	9.54	2.41
2	37	74	58	95	26.36	10.89
3	43	129	101	224	45.91	25.69
4	45	180	146	404	66.36	46.33
5	31	155	177	559	80.45	64.60
6	19	114	196	673	89.10	77.17
7	10	70	206	743	93.63	85.20
8	5	40	211	783	95.90	89.79
9	4	36	215	819	97.72	93.92
10	3	30	218	849	99.10	97.36
11	1	11	219	860	99.54	98.62
12	1	12	220	872	100.00	100.00
	$N_1=220$	$N_2=872$				

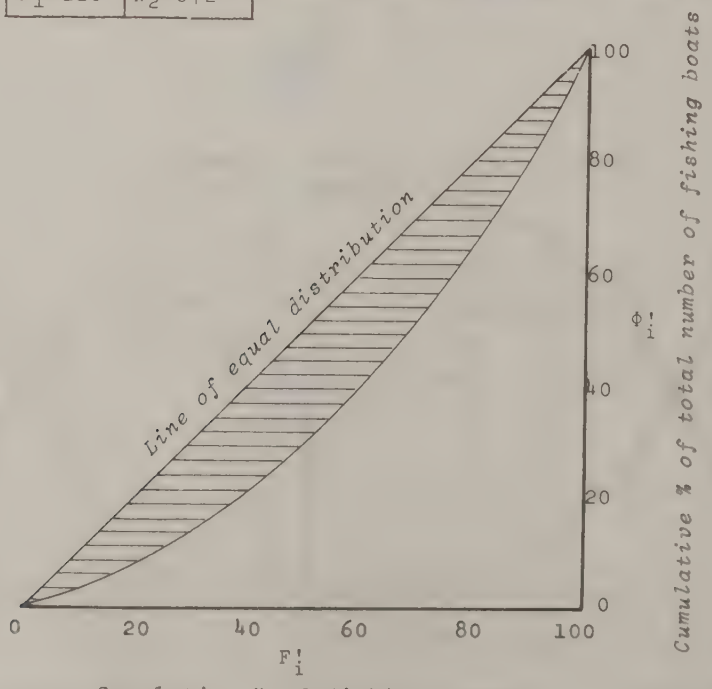


Figure 2.10.2.1 Cumulative percentage distribution of fishing camps and their fishing boats

The chart (Figure 2.10.2.1) clearly portrays the relative amount of dispersion in the frequency distribution. A curve of equal distribution would be a straight line extending diagonally from zero to 100. In our example, such a curve would indicate that any specified proportion of fishing camps would have precisely the same number of fishing boats. For example, 25 percent of the fishing camps would have 25 percent of the existing fishing boats, 50 percent of the fishing camps would have 50 percent of the fishing boats, and so on. The distance between the curve and the diagonal line is indicative of the degree of deviation.

Algebraically, the above argument can be justified as follows. Assuming that the population consists of N units, $a_1, a_2, a_3, \ldots, a_N$, and that $x_1, x_2, x_3, \ldots, x_N$, are the measurements of the units for a given characteristic, i.e.

$$\begin{pmatrix} a_1, & a_2, & a_3, & \ldots & a_i, & \ldots, & a_N \\ x_1, & x_2, & x_3, & \ldots & x_i, & \ldots, & x_N \end{pmatrix}$$

If $X = \sum_{i=1}^{N} x_i$, then, in the case of equal distribution the following equations are valid:

i) $x_1 = x_2 = x_3 = \ldots = x_N$

ii) $x_i = \frac{X}{N}$

Now, if we consider the subsets (h),

$$\begin{pmatrix} a_1, & (a_1, a_2), & (a_1, a_2, a_3) & \ldots\ldots \\ x_1, & (x_1, x_2), & (x_1, x_2, x_3) & \ldots\ldots \end{pmatrix}$$

the ratio $F_h' = \frac{h}{N}$ expresses the proportion of units in the h^{th} subset and the ratio $\phi_h' = \frac{h}{X}$ expresses the proportion of values of the characteristic in the h^{th} subset. By using the established notation the following equations are valid (values of x have been arranged in rank order),

a) $F_N' = \phi_N' = 100\%$

b) $F_h' = \phi_h'$ if $x_1 = x_2 = \ldots = x_N$ (equal distribution)

c) $F_h' > \phi_h'$ if $x_1 < x_2 \ldots < x_N$ (unequal distribution)

The size of the area E, which is confined between the curve and the line of equal distribution, indicates the level of deviation from an equal distribution. The values of E range between, (Figure 2.10.2.2),

$$0 \leq E \leq \frac{1}{2}$$

Specifically, in the case of equal distribution E = 0. If $X = \sum_{i}^{N} X_i$ corresponds to the last unit then $E = \frac{1}{2}$.

Further notes on Lorenz curve:[1/]

In the following figure (Figure 2.10.2.2) it is indicated that the area of E is given by the difference,

$$E = \frac{1}{2} - \varepsilon$$

where $\frac{1}{2}$ is the total area of the right triangle.

1/ See also, G.P. Bazigos, Sampling Techniques in Inland Fisheries with Special Reference to Volta Lake, St.S./1, FIO/SF/GHA/10, May 1970, p.6, 39.

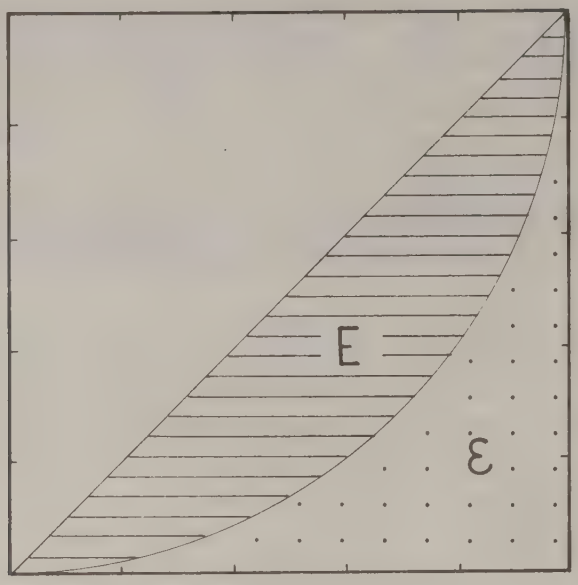


Figure 2.10.2.2

It has also been proved that:

$$E = \frac{d_1}{4m}$$

where,

d_1 : the Mean Difference of Gini (See section 4.2)

m : the mean of the distribution.

Further, the Gini's Concentration Coefficient (G_1) gives the relative deviation (%) of a given empirical distribution from the equal distribution:

$$G_1 = \frac{d_1}{2m} \times 100 = \ldots (\%), \quad (G_1 = 2E)$$

Illustration

By using the data of the above example calculate d_1, E and G_1.

$d_1 = 2.28$

$m = 3.9636$

$$E = \frac{2.28}{15.8544} = 0.1438$$

$$G_1 = \frac{2.28}{2 \times 3.9636} \times 100 = 0.2876 \text{ or } 28.76\%$$

2.11 Probability graphs

Probability paper is the name given to the special arrangement of vertical and horizontal spacing of a grid which has the property of presenting the ogive of a normal curve as a straight line. Since the normal curve is only one of a number of curves which may be fitted to a frequency distribution, the possibility of using simple graphs as a guide to tell us when to fit a normal curve, or, when fitted, if it is suitable, is of great statistical value[1/].

1/ See also Chapter 9.

The scale for the horizontal axis of probability graphs may be ruled either arithmetically or logarithmically. In the former case we have an arithmetic probability paper and in the latter case a logarithmic probability paper. The scaling of the vertical axis has been so arranged that the cumulative heights of ordinates of any perfectly normal distribution will appear as a straight line when plotted on probability paper. It should be noted that if the ogive is expressed as a straight line on arithmetic probability paper, the original frequency distribution can be considered symmetrical upon an arithmetic projection; similarly, if an ogive is presented by a straight line on logarithmic probability paper, the original frequency distribution can be considered asymmetrical, and can be symmetrical by plotting on logarithmic projection. This provides a test for proportional asymmetry.

Example

In the table below (Table 2.11.1) information is provided on the time spent by the fishing boats, harboured in a permanent fishing site, in order to get to their fishing grounds. Does the "fish searching time" distribution of fishing boats follow the normal distribution? (use arithmetic probability paper).

Table 2.11.1 Fish searching time distribution of fishing boats

Time, x (minutes)	No.of fishing boats f_i	$\sum_i f_i$	$F_i = \frac{\sum_i f_i}{N}$ (%)	Remarks
15-25	1	1	0.65	
25-35	1	2	1.31	
35-45	3	5	3.27	
45-55	13	18	11.76	
55-65	17	35	22.87	
65-75	25	60	39.21	
75-85	35	95	62.09	
85-95	25	120	78.43	
95-105	10	130	84.97	
105-115	15	145	94.77	
115-125	5	150	98.04	
125-135	2	152	99.35	
135-145	1	153	100.00	
	N=153			

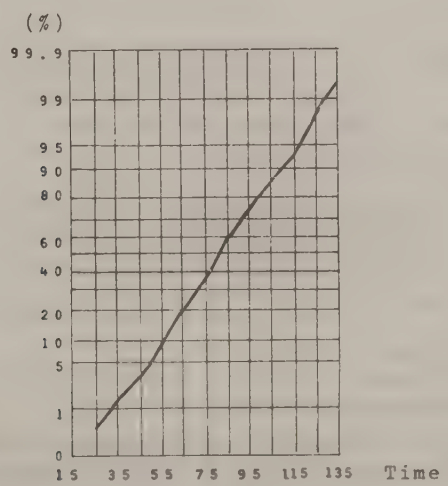


Figure 2.11.1 Cumulative distribution of fishing boats according to fish searching time, shown on arithmetic probability paper

From the above diagram one can see that the resulting curve is approximately a straight line. We can proceed with assurance to fit a normal curve to the empirical frequency distribution.

3. MEASURES OF CENTRAL TENDENCY

In this chapter we are dealing with measures of central tendency which are of use in the field of fishery statistics. Measures of central tendency are typical values which are calculated by making use of the empirical data to characterize this aspect of an empirical distribution.

3.1 The arithmetic mean (or average or mean)

Assuming that a small fishing site harbours five fishing economic units (FEU's) and that the output of the fishing operations of the units in a given day (do) is,

units:	a_1,	a_2,	a_3,	a_4,	a_5
fish catch (kg):	5,	8,	3,	9,	5

the arithmetic mean (m) of the above five numbers (simple distribution) would be given by,

$$m = \frac{1}{N}(x_1+x_2+x_3+x_4+x_5) = \frac{1}{5}(5+8+3+9+5) = \frac{30}{5} =$$

$$= 6\text{kg of fish catch per FEU/day}$$

Or, more briefly,[1/]

$$m = \frac{1}{N}\sum_i x_i, \text{ simple arithmetic mean}$$

Properties of the arithmetic mean:

a) $m = \frac{1}{N}\sum_i x_i$

and,

$$\sum_i x_i = Nm, \quad 30\text{kg} = 5\times 6\text{kg}$$

b) $\sum_i (x_i-m) = 0$

This equation indicates that the algebraic sum of the deviations of the various values of the variate x from the mean equals zero. In our example,

$$\sum_i (x_i-m) = (x_1-m)+(x_2-m)+(x_3-m)+(x_4-m)+(x_5-m)$$
$$= (-1)+(2)+(-3)+(3)+(-1) = 0$$

c) The sum of the square deviations, $\sum_i (x_i-m)^2$ is less when the deviations are taken around m than when they are taken around any other value, i.e.,

$$\sum_i (x_i-m)^2 < \sum_i (x_i-a)^2$$

where a, any other value.

1/ See, G.P. Bazigos (1974): The Design of Fisheries Statistical Surveys - Inland Waters, FAO Fish.Tech.Pap., FIPS/T133, p.89-96 Summators, expectation techniques.

In the case of frequency distributions, if the symbols f_1, f_2, f_3, etc., indicate the numbers or frequencies associated with each value being averaged, we have[1/],

$$m = \frac{1}{\sum_i f_i} \sum_i f_i x_i, \text{ weighted arithmetic mean}$$

If $\sum_i f_i = N$,

$$m = \frac{1}{N} \sum_i f_i x_i$$

Example

In the tables below (Tables 3.1.1a, b) are given the frequency distributions of 40 fishing economic units (FEU's) according to the number of traps owned and average daily catches respectively. Calculate the means of the two distributions (m_1, m_2).

Table 3.1.1a Frequency distribution of the fishing economic units according to the number of fishing traps owned

No. of traps x_i	No. of FEU's f_i	$f_i x_i$	Remarks
1	15	15	($\sum_i f_i = N = 40$)
2	10	20	
3	8	24	
4	6	24	
5	1	5	
	40	88	

$$m_1 = \frac{1}{40} \times 88 = 2.2 \text{ fish traps/FEU}$$

Table 3.1.1b Frequency distribution of the fishing economic units according to the daily average catches (x:kg)

Catches x(kg)	No.of FEU's f_i	Catches - kg x_i	$f_i x_i$	Remarks
2-6	20	4	80	($\sum_i f_i = N = 40$)
6-10	10	8	80	
10-14	6	12	72	
14-18	2	16	32	
18-22	2	20	40	
	40		304	

$$m_2 = \frac{1}{40} \times 304 = 7.6 \text{ kg of fish catch per FEU/day}$$

1/ In the case of frequency distributions with class intervals the x_i in the equation stands for the central value of the i^{th} class interval i.e. $x_i = (x_{max.} + x_{min.})/2$.

3.2 The median

The median (M) is that value of the variate which divides the empirical distribution in the whole range into two equal parts.

For a set of measurements arranged in order of magnitude, the median is defined as the middle measurement, if there is one, otherwise as the interpolated middle value.

Example

The following two simple distributions give the number of traps per fishing economic unit in two fishing sites (A, B), calculate the value of the medians respectively:

A: Number of traps per FEU x: 1, 5, 3, 8, 6, 4, 10

B: Number of traps per FEU x: 6, 12, 1, 3, 7, 14

Solution:

A: a) Number of FEU's, $N_1 = 7$

b) Rank order of the values,

x: 1, 3, 4, 5, 6, 8, 10

c) Location of the median $\frac{N_1+1}{2} = \frac{7+1}{2} = 4^{th}$ value

d) Value of the median, $M_1 = 5$ traps

B: a) Number of FEU's $N_2 = 6$

b) Rank order of the values,

x: 1, 3, 6, 7, 12, 14

c) Location of the median $\frac{N_2+1}{2} = \frac{6+1}{2} = 3.5^{th}$ value

d) Value of the median $M_2 = \frac{6+7}{2} = 6.5$ traps

The calculated value of $M_2 = 6.5$ indicates that in fishing site B one-half of the FEU's have less than 6.5 traps and one-half of the FEU's have more than 6.5 traps.

Properties of the median: An important property of the median is that it is influenced by the position of the items in the array but not by the size of the items.

The median from grouped data: The table below (Table 3.2.1) gives the frequency distribution of 100 fishing economic units according to the number of hooks owned. Calculate the median of the distribution.

Table 3.2.1 Frequency distribution of 100 fishing economic units (FEU's) according to the number of hooks owned

No.of hooks x	No. of FEU's f_i	Cumulative column F_i	Remarks
25-34	6	6	
35-44	12	18	
45-54	20	38 ← F_i	
55-64	35	73 ←	Location of median
65-74	15	88	
75-84	8	96	
85-94	4	100	
	100		

To determine the value of the median the following steps are taken:

1. Preparation of the cumulative values of f_i, (F_i)
2. Determination of the class of the frequency distribution in which the median is located. In our example the location of the median is defined,

   $\frac{N}{2} = \frac{100}{2} = 50^{th}$

   The corresponding class is 55-64 hooks.
3. The value of the median is given by,

   $M = x_{min.} + \frac{\delta}{f_i}(\frac{N}{2} - F_i)$

   where,

   x_{min}: the lower limit of the class (55 hooks)

   δ: the interval of the class (δ = 64-55 = 9 hooks)

   and,

   $M = 55 + \frac{9}{35}(\frac{100}{2} - 38) = 58.08 \simeq 58$ hooks

From the above example it is obvious that the calculated value of the median is obtained by interpolating 12 frequencies $(\frac{N}{2} - 38)$ into the fourth class, assuming that the frequencies in that class are evenly distributed within the class.

It should be noted that the computation of the median from a frequency distribution of unequal class intervals does not differ from that described above. Neither does the presence of indeterminate groups at either or both ends complicate the procedure.

Determination of the median graphically: To obtain the value of the median of a frequency distribution graphically the following steps should be taken:

1. Preparation of the ogive graph[1/].
2. Computation of,

   $\frac{\sum_i f_i}{2} = \frac{N}{2}$

   and location of the point on Y-axis.
3. Draw a perpendicular to the Y-axis at this point and extend the perpendicular to intersect the ogive.
4. At the intersection, drop a perpendicular to the X-axis. The intersection gives the value of the median.

In the following figure (Figure 3.2.1) the value of the median, located graphically, is M ≃ 58. (Data of the above example, Table 3.2.1)

1/ See section 2.10.1.

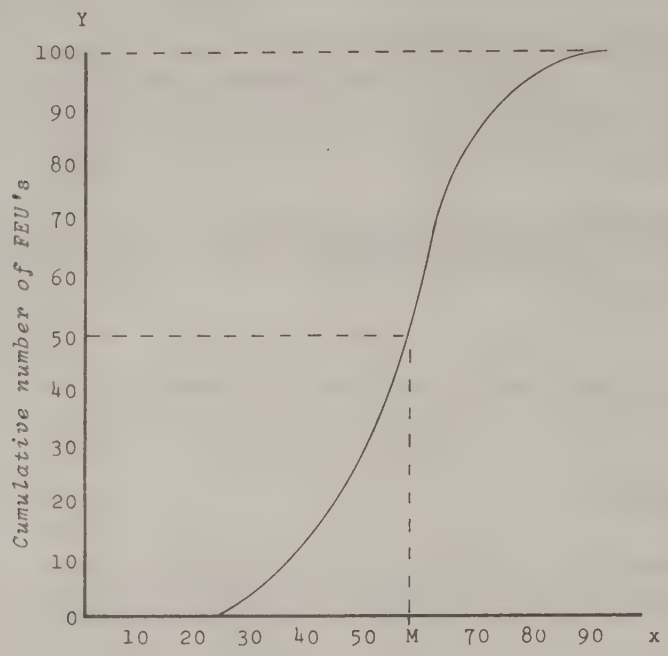


Figure 3.2.1

3.3 The quartiles, deciles

Allied measures to the median are other measures of a frequency distribution which are used to assist in measuring dispersion and skewness[1/]. We shall discuss here the quartiles and deciles.

Quartiles: There are two quartiles, the first Q_1, or lower quartile, and the third Q_3, or upper quartile. It should be noted that the three typical values Q_1, M, Q_3, divide the distribution into four equal parts (Figure 3.3.1);

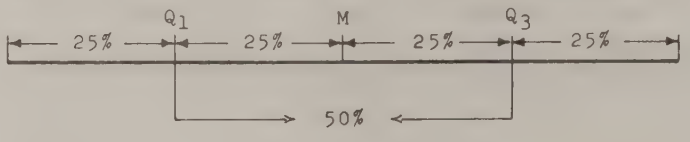


Figure 3.3.1

1/ See Chapter 4.

In the case of a simple distribution the location of the quartiles is defined, for Q_1 by $\frac{N+1}{4}$ and for Q_3 by $\frac{3(N+1)}{4}$

Example

By using the data of the above example (FEU's in fishing site A) calculate the Q_1 and Q_3 of the distribution (x: number of traps per FEU).

x: 1, 3, 4, 5, 6, 8, 10

a) Location of Q_1 : $\frac{7+1}{4} = 2^{nd}$

Value of Q_1 = 3 traps

b) Location of Q_3 : $\frac{3(7+1)}{4} = 6^{th}$

Value of Q_3 = 8 traps

In the case of a frequency distribution the locations of Q_1, Q_3 are defined by $\frac{N}{4}$ and $\frac{3N}{4}$ respectively. Their values are given by the formulae:

$$Q_1 = x_{min} + \frac{\delta}{f_i}\left(\frac{N}{4} - F_i\right)$$

$$Q_3 = x_{min} + \frac{\delta}{f_i}\left(\frac{3N}{4} - F_i\right)$$

Deciles: Like quartiles, the first decile (D_1) corresponds to the lower 10 percent and the ninth (D_9) to the upper 10 percent. In the case of simple distributions the locations of D_1 and D_9 are defined by $\frac{(N+1)}{10}$ and $\frac{9(N+1)}{10}$. In the case of frequency distributions their locations are defined by $\frac{N}{10}$ and $\frac{9N}{10}$ respectively. Their values are calculated by the formulae,

$$D_1 = x_{min} + \frac{\delta}{f_i}\left(\frac{N}{10} - F_i\right)$$

$$D_9 = x_{min} + \frac{\delta}{f_i}\left(\frac{9N}{10} - F_i\right) \quad \text{etc.}$$

Like the median, the values of quartiles and deciles can be defined graphically;

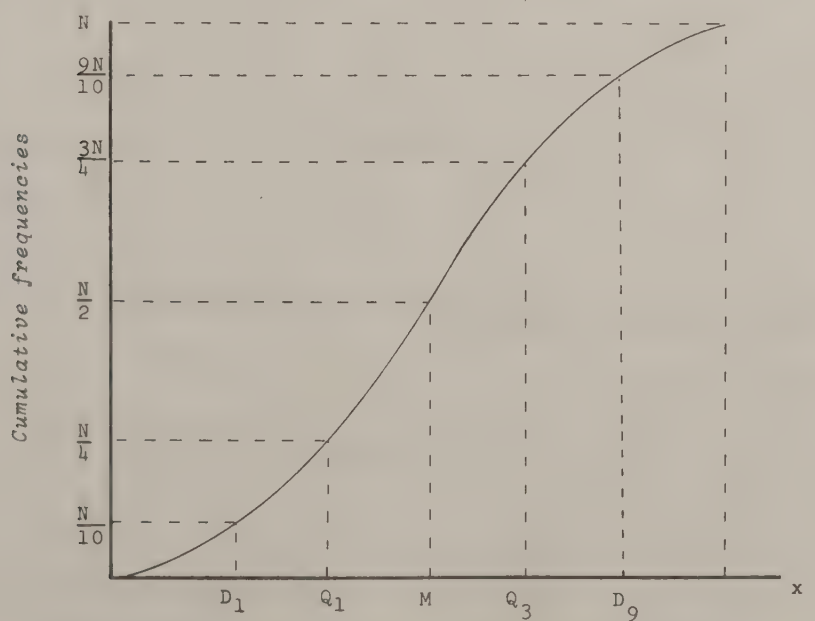


Figure 3.3.2

3.4 The mode (from grouped data)

The mode of a set of measurements is defined as the measurement with the maximum frequency, if there is one. If there is more than one measurement with the maximum frequency, no completely satisfactory definition exists. When the modal class has been defined the value of the mode (To) is given by the formula,

$$To = x_{min} + \frac{\delta \times D_1}{D_1 + D_2}$$

where,

x_{min}: The lower value of the modal class.

δ: The interval of the modal class.

D_1: The difference between the frequency of the modal class and the frequency of the preceding class (sign neglected).

D_2: The difference between the frequency of the modal class and the frequency of the following class (sign neglected).

Example

By using the data of Table 3.2.1 calculate the value of To:

1. Modal class 55-64
2. $\delta = 9$
3. $D_1 = 35$
4. $D_2 = 15$

$$To = 55 + \frac{9 \times 35}{15 + 35} = 61.3 \text{ hooks}$$

3.5 Minor means

3.5.1 Geometric mean

The geometric mean (G) of a set of measurements (N) (simple distribution) assuming they are positive, is defined by,

$$G = \sqrt[N]{x_1 x_2 x_3 \ldots x_N}$$

or,

$$\log G = \frac{1}{N}(\log x_1 + \log x_2 + \ldots + \log x_N)$$

or,

$$\log G = \frac{1}{N}\sum_i^N \log x_i$$

For a frequency distribution - x_i represents the i^{th} class mark and f_i the respective frequency - the G (weighted) is given by,

$$G = \sqrt[N]{x_1^{f_1} x_2^{f_2} \ldots x_N^{f_N}}, \quad (N = \sum_i f_i)$$

$$\log G = \frac{1}{N}(f_1 \log x_1 + f_2 \log x_2 + \ldots f_N \log x_N)$$

or,

$$\log G = \frac{1}{N}\sum_i^N f_i \log x_i$$

Properties of the geometric mean:

1. Any series of numbers having the same N and the same product have the same geometric mean.
2. The product of the ratios of the values on one side of the geometric mean to the geometric mean is equal to the product of the ratios of the geometric mean to the values on the other side of the geometric mean.

Fields of application of geometric mean:

1. Below is given the distribution of FEU's according to the fishing method used in two strata of lake A.

Stratum	1. Lusenga	2. Other fishing methods	Ratio $(\frac{2}{1})$ (%)	Ratio $(\frac{1}{2})$ (%)
Str:1	800	400	50	200
Str:2	150	300	200	50

The arithmetic mean of the two ratios of other fishing methods to Lusenga FEU's is 125 percent. Likewise the arithmetic mean of two ratios of Lusenga to other fishing methods FEU's is 125 percent. These two averages are inconsistent to each other. However, the geometric mean of each of the two pairs of ratios is $\sqrt{0.50 \times 2.00} = 1.00$ or 100 percent (consistency).

2. Probably the most frequently used application of the geometric principle has to do with the determination of the average percent of change over a period of time:

Example

The filling period of a man-made lake lasted for 10 years. Before the creation of the lake there were 1000 fishing boats. When the lake was full there were 12000 fishing boats. Calculate the average annual percent of change of fishing boats.

$$U_n = U_o(1+a)^n$$

where,

U_n = 12000 fishing boats

U_o = 1000 fishing boats

n = 10 years

a = average annual percent of change of fishing boats.

Substituting in the above equation we get,

$$12000 = 1000(1+a)^{10}$$

or,

$$12 = (1+a)^{10}$$

$$\log 12 = 10 \log (1+a)$$

or,

$$\log (1+a) = \frac{\log 12}{10}$$

or,

$$\log (1+a) = \frac{1.07918}{10} = 0.107918$$

or,

$$1+a = 1.282$$

and,

$$a = 28.2 \text{ percent}$$

4. MEASURES OF SPREAD (Range, mean deviation, variance)

Two distributions may have the same mean value, same mode and median, but differ from each other according to whether the values of the variate cluster closely around the mean or are spread widely on either side. In addition, therefore, to magnitudes of position or of central tendency - mean, mode and median - we require additional magnitudes to measure the degree to which the variate values cluster about their mean or spread from it. The table below (Table 4.1) gives information on the size distribution of the fishing sites in two limnological zones of a lake. Also, Figures 4.1a, b give the line diagrams of the respective distributions.

Table 4.1 · Size distribution of the fishing sites in two limnological zones of lake A

No. of fishing boats - x -	No. of fishing sites		Remarks
	Limn.zone-1	Limn.zone-2	
1	2	0	
2	4	0	
3	8	0	
4	12	1	
5	16	5	
6	20	19	
7	27	39	
8	32	68	
9	25	43	
10	18	17	
11	15	6	
12	10	2	
13	7	0	
14	4	0	
Total	200	200	

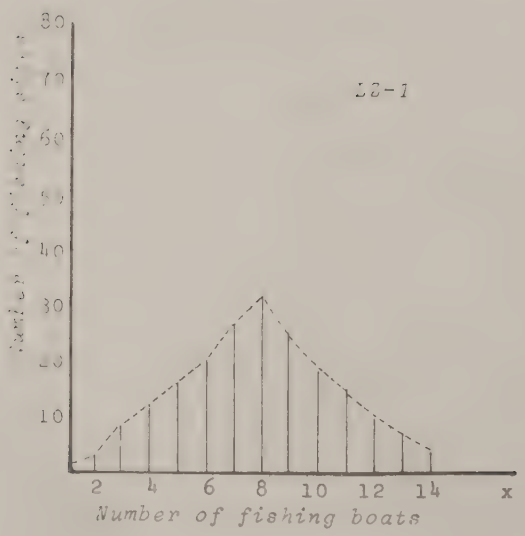


Figure 4.1a

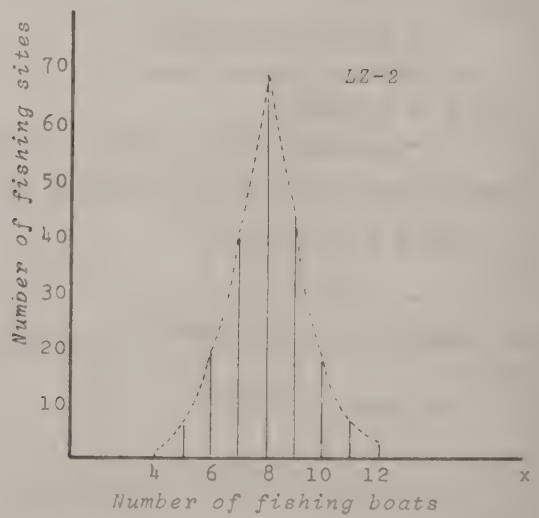


Figure 4.1b

In our example the two distributions have almost the same average number of fishing boats per fishing site. However, the level of dispersion in limnological zone 1 is bigger than that in limnological zone 2. For the measure of dispersion of an empirical distribution different magnitudes can be used i.e. a) range, b) mean deviation, and c) variance.

Range: A crude measure of the level of dispersion is given by the range (R),

$$R = x_{max.} - x_{min.}$$

However, the range has certain disadvantages. It fails to give any consideration to the arrangement of the values between the two extreme values. Furthermore, the range is misleading if either of the extreme values is an unusual occurrence. In such cases the quartile deviation[1/] (Q) and percentile range (D) are used,

1. $Q = Q_3 - Q_1$
2. $D = D_9 - D_1$

Mean deviation: This measure of dispersion is calculated in relation to the arithmetic mean,

$$AD = \frac{1}{N}\sum_{i=1}^{N} |x_i - m| \quad \text{or} \quad AD = \frac{1}{\sum_i f_i}\sum_i f_i |x_i - m|$$

The AD is of a limited value in the field of applied statistics.

Variance: In the field of statistics a more important measure of spread is a magnitude called the variance. The variance is given by,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - m)^2, \text{ simple distribution}$$

or,

$$\sigma^2 = \frac{1}{\sum_i f_i}\sum_i f_i (x_i - m)^2, \text{ frequency distribution.}$$

In words, the variance is the mean of the squared deviations from the mean of the values of the survey variate. The positive square root of the variance is called the standard deviation,

$$\sigma = +\sqrt{\sigma^2}$$

Algebraical transformation:

$$\sigma^2 = \frac{1}{\sum_i f_i}\sum_i f_i (x_i - m)^2$$

or,

$$(\sum_i f_i)\sigma^2 = \sum_i f_i (x_i - m)^2 \text{ or } N\sigma^2 = \sum f_i (x_i - m)^2, \ (N = \sum_i f_i)$$

Expanding the right-hand side of the above equation we get,

1/ In practice the semi-interquartile range (Q') is also used, $Q' = \frac{1}{2}(Q_3 - Q_1)$.

$$N\sigma^2 = \sum_i f_i(x_i^2 - 2x_i m + m^2)$$

$$= \sum_i f_i x_i^2 - 2m\sum_i f_i x_i + Nm^2$$

$$= \sum_i f_i x_i^2 - 2Nm^2 + Nm^2$$

$$= \sum_i f_i x_i^2 - Nm^2$$

and,

$$\sigma^2 = \frac{1}{N}\left(\sum_i f_i x_i^2 - Nm^2\right)$$

Further, since,

$$m = \frac{\sum_i f_i x_i}{N}$$

Substituting m in the above equation we get,

$$\sigma^2 = \frac{1}{N}\left(\sum_i f_i x_i^2 - \frac{\left(\sum_i f_i x_i\right)^2}{N}\right)$$

Coefficient of variation: The ratio,

$$CV(x) = \frac{\sigma}{m} \times 100 = \ldots\%$$

is called the coefficient of the variation and is used as a measure of relative dispersion.

Properties of the standard deviation: If $\pm\sigma$ is measured from the arithmetic mean of a normal distribution, 68.27 percent of the values of the survey variate are included; within the range $m \pm 2\sigma$, 95.45 percent are included; and within $m \pm 3\sigma$, 99.73 percent, or nearly all, of the values are included[1/].

Example

By using the data of the above example[2/] (limnological zone 2) calculate: m, σ^2, σ and CV(x):

$$m = \frac{1}{\sum_i f_i}\sum_i f_i x_i = \frac{1607}{200} = 8 \text{ fishing boats}$$

$$\sigma^2 = \frac{1}{\sum_i f_i}\left(\sum_i f_i x_i^2 - \frac{\left(\sum_i f_i x_i\right)^2}{\sum_i f_i}\right) = \frac{1}{200}\left(13285 - \frac{(1607)^2}{200}\right)$$

$$= 1.865 \text{ fishing boats}^2$$

$$\sigma = +\sqrt{\sigma^2} = +\sqrt{1.865 \text{ fishing boats}^2} = 1.37 \text{ fishing boats}$$

$$CV(x) = \frac{\sigma}{m} \times 100 = \frac{1.37}{8} \times 100 = 17.125\%$$

1/ See section 7.5.

2/ In continuous frequency distributions, x_i refers to the central values of the respective class intervals.

4.1 Tchebycheff's inequality

For any series of values, no matter how they are distributed the following inequalities are valid,

1. The proportion of the values lying within the limits $m \pm k\sigma$ (where the value of k is greater than one) will be more than $1-\frac{1}{k^2}$
2. The proportion of the values lying beyond the limits $m \pm k\sigma$ will be less than $\frac{1}{k^2}$

It has been discussed that σ is an important measure of spread of a distribution. This in turn implies that the smaller the value of σ the greater the uniformity and vice versa. To avoid this inverse relationship, a modification referred to as a measure of precision is sometimes used, especially with reference to the precision of a series of physical measurements. This measure is,

$$h^2 = \frac{1}{2\sigma^2}$$

4.2 Gini's mean difference (d_1)

Assuming that $x_1, x_2, \ldots, x_i, \ldots, x_N$ are the values of a simple distribution. If we get the absolute differences,

$$\begin{pmatrix} |x_1-x_1|, & |x_2-x_1|, & |x_3-x_1| & \ldots & |x_N-x_1| \\ |x_1-x_2|, & |x_2-x_2|, & |x_3-x_2| & \ldots & |x_N-x_2| \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ |x_1-x_N|, & |x_2-x_N|, & |x_3-x_N| & \ldots & |x_N-x_N| \end{pmatrix}$$

the resulting table contains $NN = N^2$ absolute differences. The mean value of the obtained absolute differences is a measure of spread of the distribution and is given by,

$$d_1 = \frac{1}{N^2}\sum_i\sum_j |x_i-x_j|$$

where,

suffixes $i, j = 1, 2, 3, \ldots, N$

In the above table of absolute differences, N of the values are equal to zero, i.e. $|x_1-x_1|, |x_2-x_2|, \ldots, |x_N-x_N|$. Further, the remaining set of values $N(N-1)$, can be divided into two groups having equal values i.e. $|x_1-x_2| = |x_2-x_1|$, $|x_3-x_1| = |x_1-x_3|$ etc. In such a case the following equation is valid (x values in rank order),

$$\sum_i\sum_j |x_i-x_j| = 2\sum_{i>j}\sum |x_i-x_j|$$

and,

$$d_1 = \frac{1}{N^2}2\sum_{i>j}\sum |x_i-x_j|$$

Example

The number of gillnet units used (x) in a day (do) by the fishing economic units of a fishing site are in rank order x: 5, 8, 12, 15, 17, 20, 25, 30. Calculate the value of d_1.

x:	5	8	12	15	17	20	25	30	Horizontal totals
5	0	3	7	10	12	15	20	25	92
8		0	4	7	9	12	17	22	71
12			0	3	5	8	13	18	47
15				0	2	5	10	15	32
17					0	3	8	13	24
20						0	5	10	15
25							0	5	5
30								0	
								$\sum_{i>j}\sum(x_i-x_j)$ =	286

$$d_1 = \frac{2\times 286}{(8)^2} = 8.94 \text{ gillnet units}$$

The calculated value of d_1 indicates that, on average (absolute), each value differs from each of the others by 8.94 gillnet units.

In the case of a frequency distribution the formula for d_1 is given:

$$d_1 = \frac{2\delta}{N^2}\sum_i F_i(N-F_i)$$

where, $N = \sum_i f_i$, F_i: cumulative frequencies

and δ: fixed class interval

If the class interval is not fixed, then,

$$d_1 = \frac{2}{N^2}\sum_i (x_{i+1}-x_i)F_i(N-F_i)$$

Example

In the following table the distribution of the fishing economic units of a fishing site according to the number of hooks owned (z) by the units is given. Calculate the value of d_1.

Distribution of the fishing economic units of fishing site B according to the number of hooks owned by the units

No. of hooks owned -z-	Fishing economic units f_i	F_i	$N-F_i$	$F_i(N-F_i)$	Remarks
1-10	6	6	94	564	
11-20	12	18	82	1476	
21-30	20	38	62	2356	(δ = 10)
31-40	35	73	27	1971	
41-50	15	88	12	1056	
51-60	8	96	4	384	
61-70	4	100	0	-	
	N=100			7807	

$$d_1 = \frac{2x10}{(100)^2} \times 7807 = 15.61 \text{ hooks}$$

5. MOMENTS

5.1 Moments

Moments are used in statistics for an arithmetical description of distributions. Let us consider the following simple frequency distribution,

$$\begin{pmatrix} x_1 & f_1 \\ x_2 & f_2 \\ x_3 & f_3 \\ . & . \\ . & . \\ . & . \end{pmatrix}$$

The k^{th} moment about the origin of the above empirical distribution is given by,

$$\nu_k = \frac{1}{\sum_i^h f_i} \sum_i^h f_i x_i^k$$

If $f_i = 1$ and $h = n$, then,

$$\nu_k = \frac{1}{n} \sum_i^n x_i^k$$

For k = 0, 1, 2, 3, ... the corresponding moments about the origin are,

$$\nu_0 = \frac{1}{\sum_i f_i} \sum_i f_i x_i^0 = \frac{1}{\sum_i f_i} \sum_i f_i = 1$$

Mean,

$$\nu_1 = m = \frac{1}{\sum_i f_i} \sum_i f_i x_i$$

$$\nu_2 = \frac{1}{\sum_i f_i} \sum_i f_i x_i^2$$

$$\nu_3 = \frac{1}{\sum_i f_i} \sum_i f_i x_i^3$$

etc.

The k^{th} moment about the mean of an empirical distribution is given by,

$$\mu_k = \frac{1}{\sum_i f_i} \sum_i^h f_i (x_i - m)^k$$

If $f_i = 1$ and $h = n$, then,

$$\mu_k = \frac{1}{n} \sum_i^n (x_i - m)^k$$

For k = 0, 1, 2, 3, ... the corresponding moments about the mean are,

$$\mu_0 = \frac{1}{\sum_i f_i} \sum_i f_i (x_i - m)^0 = \frac{1}{\sum_i f_i} \sum_i f_i = 1$$

$$\mu_1 = \frac{1}{\sum_i f_i} \sum_i f_i (x_i - m) = 0$$

Variance,

$$\mu_2 = \sigma^2 = \frac{1}{\sum_i f_i} \sum_i f_i (x_i - m)^2$$

etc.

5.2 Higher moments

In the field of applied statistics the following magnitudes are usually used to describe an empirical frequency distribution,

$m = \nu_1$: mean

$\sigma^2 = \mu_2$: variance

β_1 : measure of relative skewness[1/]

β_2 : measure of peakedness of the distribution (kurtosis)

Specifically, β_1 and β_2 are expressed as functions of higher moments,

---

1/ Skewness is also sometimes measured by α_3, where $\alpha_3 = \sqrt{\beta_1}$.

$$\beta_1 = \mu_3^2/\mu_2^3$$

$$\beta_2 = \mu_4/\mu_2^2$$

It should be noted that the value of β_1 is zero for a symmetrical distribution and different from zero for a non-symmetrical distribution:

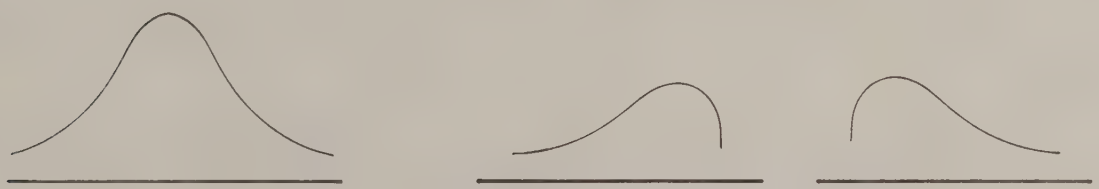


Figure 5.2.1

The magnitude β_2 has a value of 3.0 for the normal curve. For a platykurtic curve, $\beta_2<3.0$. For a leptokurtic curve, $\beta_2>3.0$.

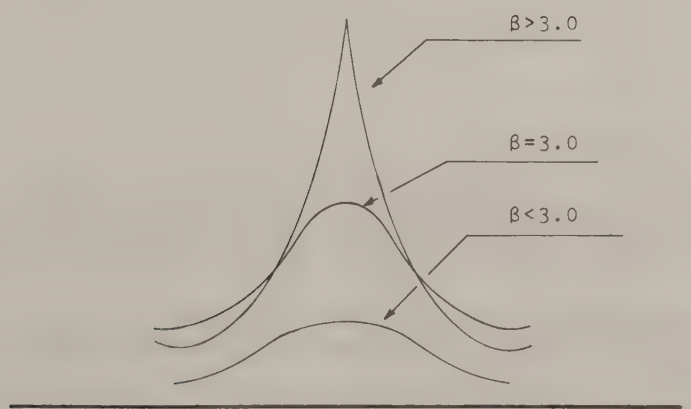


Figure 5.2.2

Relationship between μ_k and ν_k moments:

General formula:

$$\mu_k = (\nu-\nu_1)^k$$

Specifically, for k = 0, 1, 2, 3, 4, we have,

$$\mu_0 = (\nu-\nu_1)^0 = 1$$

$$\mu_1 = (\nu-\nu_1)^1 = \nu_1-\nu_1 = 0$$

$$\mu_2 = (\nu-\nu_1)^2 = \nu_2-2\nu_1\nu_1+\nu_1^2 = \nu_2-\nu_1^2$$

$$\mu_3 = (\nu-\nu_1)^3 = \nu_3-3\nu_1\nu_2+3\nu_1\nu_1^2-\nu_1^3 = \nu_3-3\nu_1\nu_2+2\nu_1^3$$

$$\mu_4 = (\nu-\nu_1)^4 = \nu_4-4\nu_1\nu_3+6\nu_1^2\nu_2-4\nu_1\nu_1^3+\nu_1^4 = \nu_4-4\nu_1\nu_3+6\nu_1^2\nu_2-3\nu_1^4$$

Example

By using available data concerning the frequency distribution of FEU's according to the number of gillnet units owned, the moments μ_2, μ_3 and μ_4 were calculated (μ_2 = 193.6, μ_3 = 244.782, μ_4 = 104205.7). Calculate β_1 and β_2.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = (244.782)^2 \div (193.6)^3 = 0.008$$

(almost symmetrical distribution)

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 104205.7 \div (193.6)^2 = 2.78$$

(slightly platykurtic curve)

5.3 Measures of skewness

To compare the skewness of two distributions, it is necessary to have some measure of skewness which will not depend upon the particular units used. Since the mode is not influenced by the presence of extreme values, the median is influenced by their position only, and the arithmetic mean is influenced by the size of the extremes, we could use the parameters to measure skewness. The Personian measures of skewness are relative measures of skewness and are given by,

1. $S_{k1} = \frac{m-To}{\sigma}$

2. $S_{k2} = \frac{3(m-M)}{\sigma}$

In the above case (1) if a frequency distribution is unimodal and the mode lies to the right of the mean, there will be a long tail on the negative side, and the curve is accordingly called negative skew, $S_{k1}<0$. If however, the mode lies to the left of the mean, the curve is called positive skew, $S_{k1}>0$.

There is, however, some difficulty in most frequency distributions to calculate accurately the value of the mode. The median may be more satisfactorily located, and therefore we use S_{k2}.

Another measure (Bowley) is given by,

$$S_{k3} = \frac{(Q_3-M)-(M-Q_1)}{(Q_3-M)+(M-Q_1)}$$

The value of S_{k3} ranges between -1 and +1.

Example

The calculated values of m, To, M, Q_1, Q_3, σ, of a frequency distribution are given below. Calculate S_{k1}, S_{k2}, S_{k3}.

$m = 58.10$

$To = 59.30$

$M = 58.43$

$Q_1 = 48.50$

$Q_3 = 66.33$

$\sigma = 13.89$

$$S_{k1} = \frac{m-To}{\sigma} = \frac{58.10-59.30}{13.89} = -0.09$$

$$S_{k2} = \frac{3(m-M)}{\sigma} = \frac{3(58.10-58.43)}{13.89} = -0.07$$

$$S_{k3} = \frac{(Q_3-M)-(M-Q_1)}{(Q_3-M)+(M-Q_1)}$$

$$= \frac{(66.33-58.43)-(58.43-48.5)}{(66.33-58.43)+(58.43-48.5)} = -0.11$$

6. PROBABILITIES IN A FINITE SAMPLE SPACE

6.1 Concepts and definitions

The statistician is interested in probability only as it pertains to the possible outcomes of experiments. Furthermore, he is interested in only those experiments that are repetitive in nature, or that can be conceived of as being so. Experiments such as tossing a coin, or reading the daily surface temperature of a lake with a thermometer are examples of simple repetitive experiments.

From a statistical point of view, a description of an experiment should cover the following aspects:

1. The purpose of the experiment.
2. The experimental conditions of the experiment i.e. how the experiment is to be carried out.
3. Description of the possible outcomes of the experiment.

Sample space; sample point: A sample space of an experiment is a set (ss) of elements such that any outcome of the experiment corresponds to exactly one element in the set. An element in a sample space is called a sample point. It should be noted that the probability that is assigned to a given sample point should be approximately equal to the proportion of times that the sample point will be obtained, or is expected to be obtained, in a large number of repetitions of the experiment.

Event (E): An event of an experiment is defined as a collection of sample points all of which possess a particular property. To obtain the probability of an event E, add the probabilities assigned to the elements of the subset (ss) that corresponds to E.

Mutually exclusive events: If two events cannot happen at the same time, they are said to be mutually exclusive. Specifically, if two events have no points in common they are called mutually exclusive, or disjoint. And n events are mutually exclusive if no two of them have any points in common.

Partition: If the events E_1, E_2 ... , E_m are mutually exclusive and exhaustive (i.e., their union contains all the sample points of (s)), then we say that the m events form a partition of the sample space (s) into m subsets.

Complementary events: An event E and the event $\bar{E}$, consisting of all points of the same sample space not in E, are called complementary events.

6.2 Applications of probability

Applications of probability are often concerned with a number of related events rather than with just one event. Further, with reference to the sample space (s), if E_1, E_2, etc. are events in (s), the probability of an event written $P(E_i)$ is a number possessing the following properties,

1. $P(E_i) \neq 0$
2. $\sum_i P(E_i) = 1$

Addition theorem: Consider two events E_1, E_2, associated with an experiment. One may be interested in knowing whether at least one of the events E_1 and E_2 will occur when the experiment is performed. This event will be denoted by the sum E_1+E_2 and the probability by $P(E_1+E_2)$. On the other hand, one may be interested in knowing whether both E_1 and E_2 will occur when the experiment is performed. This joint event will be denoted by the product E_1E_2 and the probability by $P(E_1E_2)$ (see multiplication theorem).

In the former case, when E_1 and E_2 are mutually exclusive,

$$P(E_1+E_2) = P(E_1)+P(E_2)$$

If events E_1 and E_2 have sample points in common (E_1E_2) then,

$$P(E_1+E_2) = P(E_1)+P(E_2)-P(E_1E_2)$$

Multiplication theorem: An aquarium contains 10 fish, six males (E_1) and four females (E_2). A fish is selected at random (without replacement) and its sex noted. The probability of selecting a male in the first trial is,

$$P(E_1) = \frac{6}{10}$$

The probability of selecting a female in the second trial subject to the condition that a male has been selected in the first trial is,

$$P(E_2/E_1) = \frac{4}{9}$$

Now, the probability of the joint event $E = E_1E_2$ i.e. male in the first trial and female in the second trial is,

$$P(E_1E_2) = P(E_1)P(E_2/E_1) = \frac{6}{10} \times \frac{4}{9} = \frac{24}{90}$$

The $P(E_2/E_1)$ is called conditional probability.

If $P(E_1) = P(E_2/E_1)$, then the event E_2 is said to be independent in a probability sense, or more briefly, independent, of the event E_1. When E_2 is independent of E_1 then,

$$P(E_1E_2) = P(E_1)P(E_2)$$

6.2.1 Illustrations

In a fish pond there are 200 fish, out of which 80 are males (E_1) and 120 are females (E_2). If sampling (random) is with replacement, what is the probability of drawing two fish both of which are female?

$$P(E_2E_2) = (\frac{120}{200})(\frac{120}{200}) = \frac{14400}{40000} = 0.36$$

In the above example what is the probability of both of them being males?

$$P(E_1E_1) = (\frac{80}{200})(\frac{80}{200}) = \frac{6400}{40000} = 0.16$$

In the above example what is the probability that the first fish will be male and the second female?

$$P(E_1E_2) = (\frac{80}{200})(\frac{120}{200}) = \frac{9600}{40000} = 0.24$$

If sampling is made without replacement what is the probability: a) The first selected fish will be female and the second selected female? b) The first selected fish will be male and the second selected male? and c) The first selected fish will be male and the second selected female?

Solution

a) $P(E_2, E_2/E_2) = (\frac{120}{200})(\frac{119}{199}) = \frac{14280}{39800} = 0.3588$

b) $P(E_1, E_1/E_1) = (\frac{80}{200})(\frac{79}{199}) = \frac{6320}{39800} = 0.1588$

c) $P(E_1, E_2/E_1) = (\frac{80}{200})(\frac{120}{199}) = \frac{9600}{39800} = 0.2412$

6.3 Combinatorial formulae

1. Factorial n (n!): The number of ways of arranging n items in a row is factorial n, $n!n(n-1)(n-2)...2.1$
   $5!5\times4\times3\times2\times1 = 120$

2. Permutations: The number of ways of selecting r items from n items and arranging them in a row is referred to as the number of permutations of n items taking r at a time,

   $$_nP_r = \frac{n!}{(n-r)!}$$

   If n = 4 and r = 2,

   $$_4P_2 = \frac{4!}{2!} = \frac{4\times3\times2\times1}{2\times1} = 12$$

3. Combination: The number of ways of selecting r items from n without taking into account different arrangements is referred to as the number of combinations of n items taking r at a time,

   $$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

   If n = 4 and r = 2,

   $$\binom{4}{2} = \frac{4!}{2!2!} = \frac{4\times3\times2\times1}{(1\times2)(1\times2)} = 6$$

6.3.1 Illustrations of the use of combinatorial formulae

An aquarium contains 10 fish, six males (E_1) and four females (E_2). A sample of three fish is taken at random, without replacement. Calculate the probability of the following events:

1. The sample contains three males,

   $$P(E_1E_1E_1) = \binom{6}{3} \div \binom{10}{3} = \frac{20}{120} = \frac{1}{6}$$

2. The sample contains two males and 1 female,

   $$P(E_1E_1E_2) = \frac{\binom{6}{2}\binom{4}{1}}{\binom{10}{3}} = \frac{60}{120} = \frac{1}{2}$$

3. The sample contains at least two males: This statement is satisfied when the sample contains two or three males. These events are mutually exclusive and the joint probability is given by, $P(E) = \frac{1}{6}+\frac{1}{2} = \frac{2}{3}$

7. THEORETICAL FREQUENCY DISTRIBUTIONS OF ONE VARIABLE

7.1 Random variables and their probability functions[1/]

Assuming that a fish pond contains 200 fish altogether, out of which 20 are of species A, 80 of species B and 100 of species C, we may code the species x : 1, 2, 3, respectively and find the probabilities 0.1, 0.4, and 0.5. A numerical-valued variable x, as this in the example, defined on a sample space is called a random variable.

If we tabulate the results of the above example we get,

$$\begin{Bmatrix} x: & 1, & 2, & 3 \\ f(x): & .1, & .4, & .5 \end{Bmatrix} \text{ or } \begin{Bmatrix} x: & x_1, & x_2, & x_3 \\ f(x): & f(x_1), & f(x_2), & f(x_3) \end{Bmatrix}$$

In the above tabulated data the f(x) is called the frequency function of the random variable x; f(x) yields the probability that the random variable x will assume any particular value in its range.

Probability distribution: Let X be a random variable with possible values x_1, x_2, ... , x_k and associated probabilities $f(x_1)$, $f(x_2)$, ... , $f(x_k)$, the set,

$$X: \begin{Bmatrix} x_1 & f(x_1) \\ x_2 & f(x_2) \\ . & . \\ x_k & f(x_k) \end{Bmatrix} \quad \text{or} \quad X:(x_i, f(x_i)) \quad \text{where } i = 1, 2, \ldots, k$$

is called the probability distribution of X.

7.2 Binomial distribution (discrete variable)[2/]

Some experiments are composed of repetitions of independent trials, each with two possible outcomes, E_1, E_2. The binomial probability distribution may describe the variation that occurs from one set of trials of such a binomial experiment to another. Let p be the probability that the event E_1 will occur when the experiment is performed (success). $P(E_1) = p$ and $q = 1-p$ denote the probability it will fail, $P(E_2) = q = 1-p$, $(q+p) = 1$. If n independent trials be made and x be the number of successes in the n trials of the experiment, then a formula is needed for calculating the probability of obtaining x successes in n independent trials of an event for which p is the probability of success in a single trial. The respective function is known as binomial or Bernoule frequency function,

$$f(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

The probability series of the binomial distribution is given by the expansion,

$$(q+p)^n = \sum_{o}^{n} \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

where,

$$q+p = 1$$

Moments:

Mean: $m = np$

1/ See: summators, expectation techniques, p.89-100, FAO Fish.Tech.Pap., FIPS/T133 by G.P. Bazigos, May 1974.

2/ See also section 10.2.

Variance: $\sigma^2 = npq$

$\mu_3 = npq(q-p)$

$\mu_4 = 3p^2q^2n^2+npq(1-6pq)$

$\beta_1 = \frac{(q-p)^2}{npq}$, if $p = q = \frac{1}{2}$ then $\beta_1 = 0$ (symmetry)

$\beta_2 = 3+\frac{1-6pq}{npq}$

Extensive tables[1/] are available for the binomial distributions. Usually the available tables are of two kinds: a) tables giving b(x : n, p) i.e. giving the probability of observing exactly x successes in a binomial experiment composed of n trials, and b) tables giving the probability of observing r or more successes i.e. the cumulative probability from r through n.

Example

Fish products are packed in containers each containing six pieces of the product. A sample of 500 containers was selected and examined about the processing quality of the pieces. The table below gives the obtained results. a) Estimate the moments of the theoretical distribution which describes the population from which the sample has been selected (positive binomial), b) Estimate the probability, P(x = 2), P(x = 5), P(x≤3), P(x>4) (x : stands for the number of low processing quality pieces in the containers).

x :	0	1	2	3	4	5	6	
f_i :	85	100	200	80	20	10	5	= 500 containers

Solution

$\bar{x} = \frac{1}{\sum_i f_i}\sum f_i x_i = \frac{900}{500} = 1.8$

$\bar{x} = np$ and $p = \frac{1.8}{6} = 0.30$

$\sigma^2 = npq = 6\times0.3\times0.7 = 1.26$

$\sigma = +\sqrt{1.26} = 1.1225$

$\beta_1 = \frac{(q-p)^2}{npq} = \frac{(0.7-0.3)^2}{1.26} = 0.12698$

$\beta_2 = 3+\frac{1-6pq}{npq} = 3+\frac{1-6\times0.3\times0.7}{1.26} = 2.7937$

$P(x = 2) = 0.32414$

$P(x=5) = 0.01027$

$P(x\leq3) = P(x=0)+P(x=1)+P(x=2)+P(x=3) = 0.92956$

$P(x>4) = P(x=5)+P(x=6) = 0.01103$

or, $P(x>4) = 1-P(x\leq3) = 0.01103$

1/ Biometrical Tables for Statisticians, Cabr.Un.Prs., 1954.

When n is large, the binomial distribution can be adjusted so that it is closely approximated by the standard normal distribution. It has been proved that if x represents the number of successes in n independent trials of an event for which p is the probability of success in a single trial, then the variable,

$$t = \frac{x-np}{\sqrt{npq}}$$

has a distribution that approaches the normal distribution with mean 0 and standard deviation 1 as the number of trials becomes increasingly large.

Example

In the above example calculate $P(x \leq 3)$ by using the tabulated values (areas) of the standard normal distribution:

$$t = \frac{x-np}{\sqrt{npq}}$$

In the present case we use the area to the left of x = 3; to do this in the standard normal table we need to take as the right-hand boundary for x[1/] the value $3\frac{1}{2}$,

$$t = \frac{3.5-1.8}{1.1225} = 1.514$$

$P(x \leq 3) = P(t \leq 1.514) = 0.9350$ (approximation)

7.3 Poisson distribution (discrete variable)[2/]

It has been proved that for n large and p small an approximation to the binomial frequency function is known as the Poisson frequency function and is defined by,

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

The parameter λ is the mean of the distribution ($\lambda = m$).

Moments:

Mean: $\lambda = m$

Variance: $\sigma^2 = \lambda$, $\sigma = +\sqrt{\lambda}$

$\mu_4 = 3\lambda^2+\lambda$

$\beta_1 = \frac{1}{\lambda}$

$\beta_2 = 3+\frac{1}{\lambda}$

As in the case of binomial distribution there are tables available for the Poisson distribution giving the probabilities of observing exactly a value of x and the cumulative probability from r through n.

Example

The estimated sample mean of a sample selected from a population following the Poisson distribution is $\bar{x} = 2.6$, calculate: a) $P(x = 0)$, $P(x = 1)$, $P(x \leq 2)$, b) σ, β_1, β_2.

1/ If we use the area to the right of x, the left-hand boundary of x is $x-\frac{1}{2}$.

2/ See also section 10.3.

$$P(x=0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-2.6} = 0.074274$$

$$P(x=1) = 0.193112$$

$$P(x\leq 2) = P(x=0)+P(x=1)+P(x=2) = 0.518432$$

$$\hat{\sigma} = +\sqrt{2.6} = 1.612$$

$$\beta_1 = \frac{1}{2.6} = 0.3846$$

$$\beta_2 = 3+\frac{1}{2.6} = 3.3846$$

The Poisson distribution tends to normality as $m = \lambda$ increases, as in the case of binomial distribution one can make use of the tabulated values of the standard normal distribution for the calculation of the probabilities,

$$t = \frac{x-\lambda}{\sqrt{\lambda}}$$

7.4 Negative binomial distribution (discrete variable)[1/]

It has been seen that the assumption inherent in the Poisson series is that the probability of an event remains constant, which in practice is rarely true. Any variation in the expectation of an event - in particular, any tendency for an event to increase the probability of another - will increase the variance of the distribution relative to the mean, and a negative binomial distribution will invariably better describe the data.

The probability series of the negative binomial distribution is given by the expansion,

$$(q-p)^{-k} = \frac{1}{(q-p)^k}$$

where,

$q-p = 1$ and $q = 1+p$

The mean: $\mu = kp$

The variance: $\sigma^2 = kpq = \mu q = \mu(1+p) = \mu(1+\frac{\mu}{k}) = \mu+\frac{\mu^2}{k}$

The individual terms of $(q-p)^{-k}$ are given by,

$$P(x) = (1+\frac{\mu}{k})^{-k}\frac{(k+x-1)!}{x!(k-1)!}(\frac{\mu}{\mu+k})^x$$

The reciprocal of the exponent k i.e. $\frac{1}{k}$ is a measure of the excess of variance or clumping of individuals in the population, specifically:

1. If $k\to\infty$, $\frac{1}{k}\to 0$, the distribution converges to the Poisson series i.e. $\sigma^2 = \mu$.
2. If clumping increases $k\to 0$ and $\frac{1}{k}\to\infty$ and the distribution converges to the logarithmic series.

Estimates of the parameters μ and k are calculated from the frequency distribution of the sample and are given by the statistics $\bar{x}$ and $\hat{k}$. Specifically, $\bar{x}$ is calculated by the usual way. If $s^2 = \hat{\sigma}^2$, we have,

---

1/ See also, Ellio, J.M., (1971) Some methods for statistical analysis of samples of Benthis Inverterbrates, FRES,Wat.Biol.Ass.25.

$$\sigma^2 = \mu + \frac{\mu^2}{k}$$

Substituting in the above equation the statistics for the parameters, we get,

$$s^2 = \bar{x} + \frac{\bar{x}^2}{k}$$

and,

$$\hat{k} = \frac{\bar{x}^2}{s^2 - \bar{x}} \quad \text{(approximation)}$$

Example

The frequency distribution below gives the area distribution of pelagic fish in a lake as estimated by an echo-survey (sample data). Specifically, the variate x expresses integrator readings (mm) per elementary sampling distance unit (two nautical miles) and f stands for the frequencies of distance units for each x value. By using the sample data estimate k.

x:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
f:	4	14	18	24	20	12	14	12	10	8	6	4	4	2	2	2	2

Solution

$$n = \sum_i f_i = 158$$

$$\bar{x} = \sum_i f_i x_i / n = \frac{850}{158} = 5.3797$$

$$s^2 = \frac{1}{n-1}\left\{\sum f_i x_i^2 - \frac{\left(\sum_i f_i x_i\right)^2}{n}\right\} = \frac{1}{157}\left\{6654 - \frac{(850)^2}{158}\right\}$$

$$s^2 = 13.2561$$

$$\hat{k} = \frac{\bar{x}^2}{s^2 - \bar{x}} = \frac{(5.3797)^2}{13.2561 - 5.3797} = 3.67441$$

An estimate of p is given by,

$$\hat{p} = \frac{1}{1 + \bar{x}/k} = \frac{1}{1 + \frac{5.3797}{3.67441}} = 0.4058$$

Tables[1/] give the expected probabilities for individual terms of negative binomial distribution.

7.5 Normal distribution (continuous variable)

7.5.1 What is the normal distribution?

Let us assume that the fish catch of a given research boat consists of 2000 fish of a given species A. The survey variate is the continuous variable, length per fish. In the following table (Table 7.5.1.1) the length distribution of the fish catch is given,

---

1/ Williamson, E. and Bretherton, M.H. (1963): Tables of the Negative Binomial Probability Distribution, New York.

Table 7.5.1.1 Length distribution of the fish catch

Length (x) - cm -	No. of fish	(%)	Cumulative (%)	Remarks
23-24	20	1.00	1.00	
24-25	30	1.50	2.50	
25-26	40	2.00	4.50	
26-27	120	6.00	10.50	
27-28	170	8.50	19.00	
28-29	220	11.00	30.00	
29-30	270	13.50	43.50	
30-31	300	15.00	58.50	
31-32	250	12.50	71.00	
32-33	220	11.00	82.00	
33-34	140	7.00	89.00	
34-35	100	5.00	94.00	
35-36	50	2.50	96.50	
36-37	40	2.00	98.50	
37-38	30	1.50	100.00	
Total	2000	100.00		

The figures below (Figures 7.5.1.1a, 7.5.1.1b) provide graphical presentations of the length distribution and the relative length distribution of the fish catch.1/

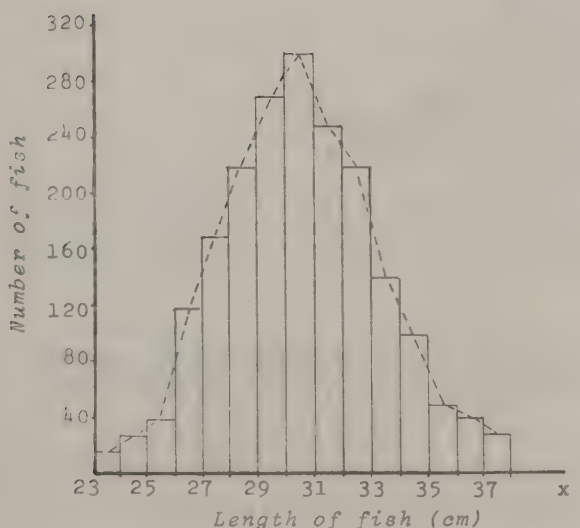


Figure 7.5.1.1a

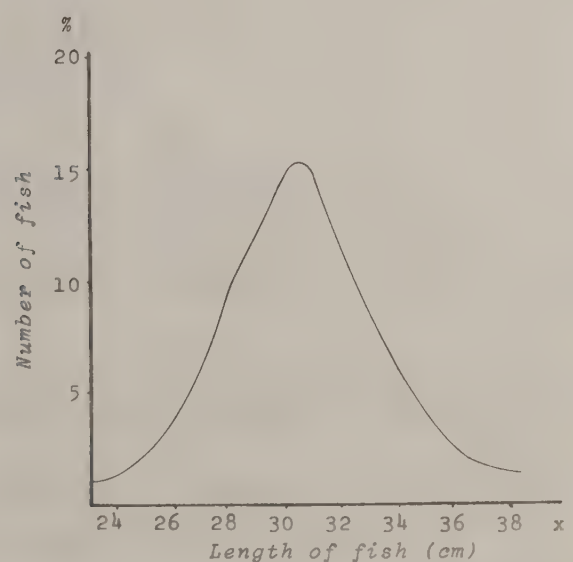


Figure 7.5.1.1b

1/ See also section 10.1.

An assessment of the produced histogram (curve) indicates that our variate varies continuously, and that as the class interval of length reduces indefinitely we come to the idea of a continuous distribution, where the relative frequency of the variate varies continuously as the variate itself varies continuously over its range. The distribution is quite symmetrical, dying out rather quickly at the tails, and possesses a shape much like that of a bell (normal distribution). Experience has shown that many variates in the field of fishery statistics and biology are normally distributed[1/].

7.5.2 Properties of the normal curve

From Figure 7.5.1.1b it is obvious that the relative frequencies associated with the length class intervals are assigned by means of the curve of the diagram, called the normal curve. From the inspection of the graph (Figure 7.5.1.1b) we note that the curve:

1. Has its highest point at x = m. The value of the ordinate at x = m equals $1/\sigma\sqrt{2\pi}$ (maximum ordinate).
2. Is symmetric about the maximum ordinate.
3. Extends without limit to the left and the right and approaches the x-axis very rapidly as we move away from x = m in either direction.

Further, it has been proved that the area under the curve and above the x-axis equals 1. The figure below (Figure 7.5.2.1) indicates some particular areas enclosed by the curve.

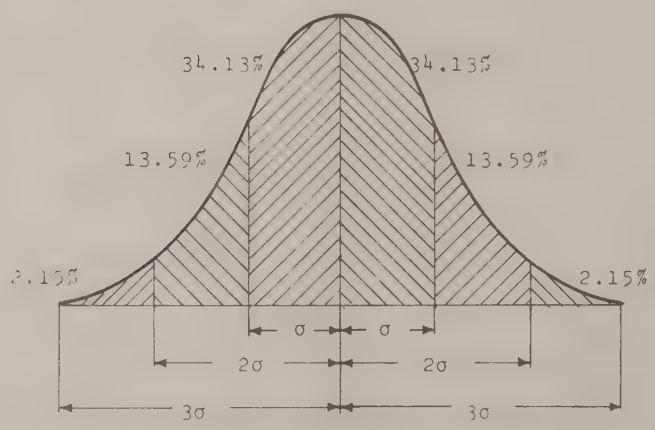


Figure 7.5.2.1

1/ A mathematical model that has proved very useful for distributions such as these is a distribution called the normal or Gaussian distribution. It is defined as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} = ce^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

Where m and σ and c are parameters, (σ>0), that make f(x) a frequency function. From the above function it is obvious that a normal distribution is completely determined by specifying its mean (m) and standard deviation (σ).

Specifically, in Figure 7.5.2.1 the area enclosed by the curve above any interval represents the probability of the variate lying in this interval.

7.5.3 Parameters of the distribution

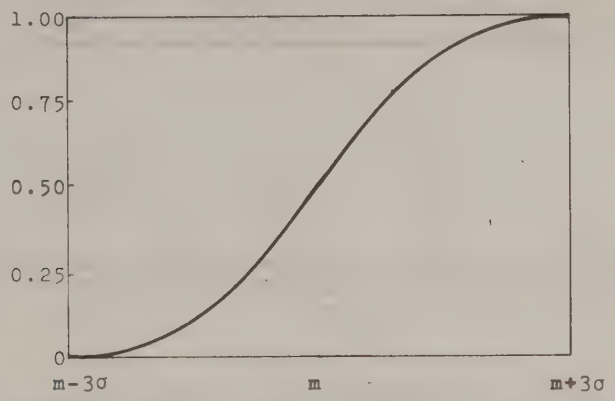
$\mu_3 = 0$

$\mu_4 = 3(\sigma^2)^2$

$\beta_1 = 0$

$\beta_2 = 3$

7.5.4 Cumulative curve



7.5.5 Sum and difference of normal variates

If x_1, x_2 are normally distributed with means m_1, m_2 and variances σ_1^2, σ_2^2, then z_1 and z_2 are normally distributed where:

$z_1 = x_1+x_2$, has mean m_1+m_2, variance $\sigma_1^2+\sigma_2^2$

$z_2 = x_1-x_2$, has mean m_1-m_2, variance $\sigma_1^2+\sigma_2^2$

7.5.6 Standard normal distribution[1/]

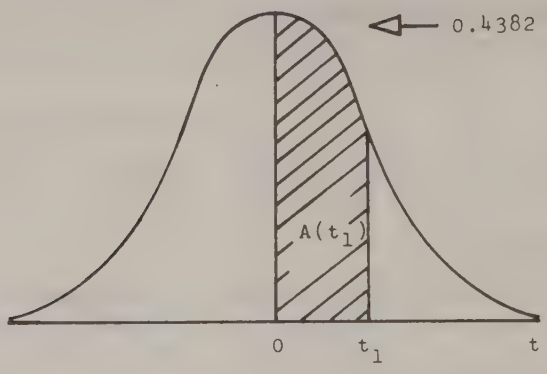
The variate $t = (x-m)/\sigma$ is called the standard normal random variable. It has been proved that the variable t is normally distributed with a mean equal to zero and a variance (standard deviation) equal to one. Tables are available (see Appendix II) giving: a) The coordinates of points on the normal standard curve for values of t between -4.00 and +4.00, at intervals of 0.1, and b) The area under the standard normal curve from 0 to t_i.

Example

If t is the standard normal variate and $t_1 = 1.54$ find the area ($A(t_1)$) under the standard normal curve from 0 to t_1. $A(t_1) = P(0<t<1.54) = .4382$.

1/ The mathematical model for the standard normal curve is:

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$



Example

The length distribution of pelagic fish at a given lake follows the normal distribution with mean length m = 5cm and standard deviation σ = 1.5cm. Find the proportion of fish between x_1 = 3cm and x_2 = 8cm.

Solution:

a) For x_1 and x_2 the respective values of the standard normal variate t are:

$t_1 = (3-5)/1.5 = -1.33$(approx), $t_2 = (8-5)/1.5 = 2$

b) $A(-1.33) = A(1.33) = 0.4082$

$A(2) = 0.4772$

c) $P(3cm<x<8cm) = A(-1.33)+A(2) = 0.8854$ or 88.54%

In the above example find the proportion of fish with length more than x_3 = 9cm.

$$t_3 = \frac{9-5}{1.5} = 2.06 \text{ (approx)}$$

$A'(2.06) = 0.50-A(2.06) = 0.50-0.4803 = 0.0197$

or,

$P(x>9cm) = 1.97\%$

In the above example find the proportion of fish with length less than x_4 = 3.5cm.

$$t_4 = \frac{3.5-5}{1.5} = -1$$

$A'(-1) = 0.50-A(1) = 0.50-0.3413 = 0.1587$

or,

$P(x<3.5cm) = 15.87\%$

8. ESTIMATION AND TESTING HYPOTHESIS

8.1 Estimation

8.1.1 The statistical definition of a random sample

From a statistical point of view the organic structure of a random sample satisfies the following two conditions:

a) $f_i(x) = f(x)$

b) $S_n(x_1, x_2 \ldots , x_n) = f(X_1).\ f(x_2) \ldots f(n)$

The meaning of condition (a): Assuming that through a given sampling method we have selected n units from an infinite population and that $x_1, x_2, \ldots , x_n$ are the population values of a given variate, by using the same experimental conditions an infinite number of samples of size n can be selected. The table below gives the order of the selected samples:

1st sample:	x_1,	x_2,	x_3,	$\ldots$,	x_i, $\ldots$,	x_n
2nd sample:	x_1',	x_2',	x_3',	$\ldots$,	x_i', $\ldots$,	x_n'
3rd sample:	x_1'',	x_2'',	x_3'',	$\ldots$,	x_i'', $\ldots$,	x_n''
.	.	.	.		.	.
.	.	.	.		.	.
.	.	.	.		.	.

In the above table the set of values of rank i (x_i, x_i', x_i'' ...) can be considered as the values of a random variate with a frequency function $f_i(x)$. If f(x) is the frequency function of the parent population, the sampling procedure produces random samples if,

$$f_1(x) = f_2(x) = f_3(x) = \ldots = f_i(x) = \ldots = f_n(x) = f(x)$$

In such a case the set of the selected n values $x_1, x_2, x_3, \ldots , x_i, \ldots , x_n$ defines a random sample of size n selected from the surveyed population.

The meaning of condition (b): Indicates the independency of the sample values.

8.1.2 Estimators

Most of the problems of estimation in statistics are problems of estimating parameters of frequency functions. For example, the Fishery Statisticians are interested in estimating the average weight of fish of species A, or the variance of the frequency function or other parameters. In a case where the value of a parameter θ in which we are interested is not known, a point-estimate of θ can be obtained by using the values $x_1, x_2, \ldots , x_n$ of a random sample selected from the surveyed population. Let $x_1, x_2, x_3, \ldots , x_n$ denote the values of a random sample. Then the function given by,

$$t = \phi(x_1, x_2, \ldots , x_n)$$

defines a function of the random variables $x_1, x_2, \ldots , x_n$ and is called the estimator of θ. For example, an estimate of the population m is given by the sample mean $\bar{x}$. In such a case the above function is,

$$t = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

It should be noted that it is customary for statisticians to use the word estimator for the function and the word estimate for the value of the function after the observational values have been inserted.

8.1.3 Sampling distribution of the estimator t

By repeating a given sampling method indefinitely we can produce a large number of samples of size n. Within each sample the value of the random variate t can be calculated. The frequency distribution of t of all possible samples of n is called the sampling distribution of the estimator t.

8.1.4 Properties of the estimators

1. The estimator t is called unbiased if $E(t) = \theta$.

   Note: the symbol E stands for expectation.

2. For a given parameter θ more than one estimator can be considered, e.g.

   $$t_1 = \phi_1(x_1, x_2, \ldots, x_n)$$
   $$t_2 = \phi_2(x_1, x_2, \ldots, x_n)$$

   If $V(t_1) < V(t_2)$, symbol V stands for variance, then estimator t_1 is called more consistent than t_2.

3. An optimum estimator is the more consistent among a group of unbiased estimators.

8.1.5 The estimator $\bar{x} = \frac{1}{n}\sum_i x_i$

It has been proved that:

1. $E(\bar{x}) = m$ (unbiased)
2. $V(\bar{x}) = \frac{\sigma^2}{n}$, $(\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}})$

Also, it has been proved that the distribution of $\bar{x}$, when the population sampled is normal $x:N(m, \sigma^2)$, is normal $\bar{x}:N(m, \frac{\sigma^2}{n})$. If the parent population is not normal then the distribution of $\bar{x}$ diverts from normality. However, the distribution of $\bar{x}$ tends to normality as the size of the sample n increases indefinitely (central limit theorem). In practice when $n>50$ the distribution of $\bar{x}$ follows the normal distribution.

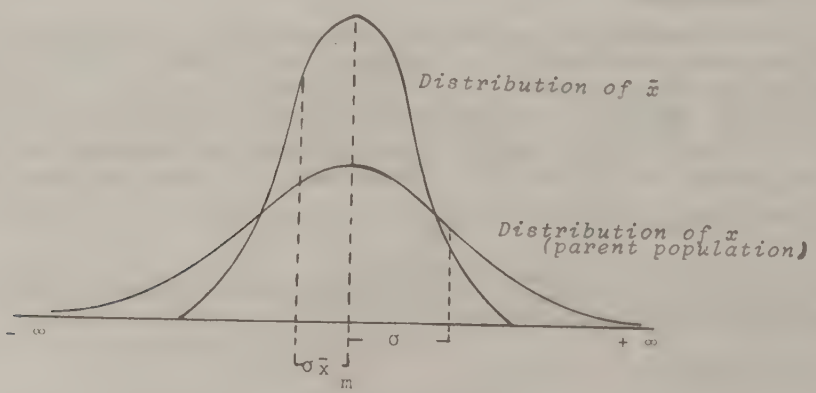


Figure 8.1.5.1

8.2 Confidence interval for m, (n>50)

It has been indicated that $\bar{x}$ is an unbiased estimator of m and that for n>50 (irrespective of the distribution of the parent population) the estimator has a normal distribution $\bar{x}:N(m,\frac{\sigma^2}{n})$. In many problems of estimation, however, one prefers an interval estimate which will express the accuracy of the estimate as well as yielding a point estimate. For estimators normally distributed the normal curve can be used to find such an interval.

The determination of the confidence interval for m by using the results of a particular sample is as follows:

1. By using the sample data we calculate the value of $\bar{x}$.
2. Although m is unknown, it is known from the statistical theory that $\bar{x}$ will be normally distributed about the m with a standard error of $\sigma_{\bar{x}}$, (σ = standard deviation of the population).
3. The fixed but unknown interval:

   $m \pm 1.96\sigma_{\bar{x}}$

   will contain 95 percent of such sample means in the long run. Further, the interval:

   $\bar{x} \pm 1.96\sigma_{\bar{x}}$

   is a variable interval, changing with each sample of n. In repeated sampling 95 percent of such intervals in the long run will contain m.

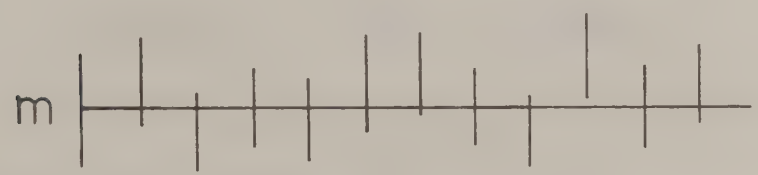


Figure 8.2.1

4. In practice, only one sample is selected and thus only one such $\bar{x}$ is available. On the basis of this one sample, the claim will be made that the interval $\bar{x} \pm 1.96\sigma_{\bar{x}}$ will contain the population mean m (P = 95 percent).

Example

To get an estimate of the average weight of fish-baskets of a given size at a lakeside market a random sample of 64 baskets full of fish was selected. Estimate the confidence interval for m (P = 95 percent), when, $\bar{x}$ = 20kg and $\sigma^2 = 9kg^2$.

Solution:

$$n = 64,\ \bar{x} = 20kg,\ \sigma_{\bar{x}} = \frac{3}{\sqrt{64}} = 0.375kg$$

$$20kg - 1.96 \times 0.375 < m < 20kg + 1.96 \times 0.375$$

$$19.265kg < m < 20.735kg$$

8.3 Testing the hypothesis[1/], Ho : m = m_o

Suppose that a biologist is certain from his experimental work that the average length of a harvestable species is m_o. Let us also assume that the standard deviation of the survey variate is known (σ). The assertion made by the biologist on the value of the unknown parameter m, is a "statistical hypothesis" and it is denoted by Ho : $m = m_o$.

A test of a statistical hypothesis is a procedure for deciding whether to accept or reject the hypothesis. In order to test the Ho the following steps should be taken:

1. A random sample of size n (n>50) is selected from the survey population. From the selected sample we calculate the value of the statistic $\bar{x}$.
2. We estimate the confidence interval for m (e.g. P = 95 percent) $\bar{x}-1.96\sigma_{\bar{x}}<m<\bar{x}+1.96\sigma_{\bar{x}}$
3. We accept the hypothesis Ho, if m_o is included in the estimated interval and vice versa. In the present case the hypothesis was tested at a confidence interval of 95 percent or level of significance a = 1-0.95 = 0.05, or 5 percent.

Example

The estimated average length of a given harvestable species is $\bar{x}$ = 174.5mm. The value of $\bar{x}$ was calculated by using the results of a random sample of size n = 100 fish. If $\sigma^2 = 6.25$, test the hypothesis that m = 176mm (a = 0.05).

Solution:

1. $n = 100$, $\bar{x} = 174.5$, $\sigma_{\bar{x}} = \frac{\sqrt{6.25}}{\sqrt{100}} = 0.25$mm
2. Confidence interval for m, (a = 0.05)
   $174.5\text{mm}-1.96\times0.25\text{mm}<m<174.5\text{mm}+1.96\times0.25\text{mm}$
   $174.01\text{mm}<m<174.99\text{mm}$
3. The H_o: m = 176mm is rejected at a = 0.05

8.4 Student's t variate

Let u be normally distributed with 0 mean and unit variance N(0, 1), let v^2 have a χ^2 with $\nu = n-1$ degrees of freedom[2/] and let u and v be independently distributed. Then the ratio,

$$t = \frac{u}{\sqrt{v^2/\nu}} = \frac{u\sqrt{n-1}}{v}$$

has Student's t distribution with ν degrees of freedom. This distribution has many similarities with the standard normal distribution (see figure following). Specifically, in practice when $\nu>60$ this distribution is considered as identical to standard normal distribution.

1/ In testing Ho, two types of error might occur: The size of type I error is the probability that the sample point will fall in the critical region when Ho is true; the size of the type II error is the probability that the sample point will fall in the non-critical region when Ho is not true.

2/ See Chapter 10.

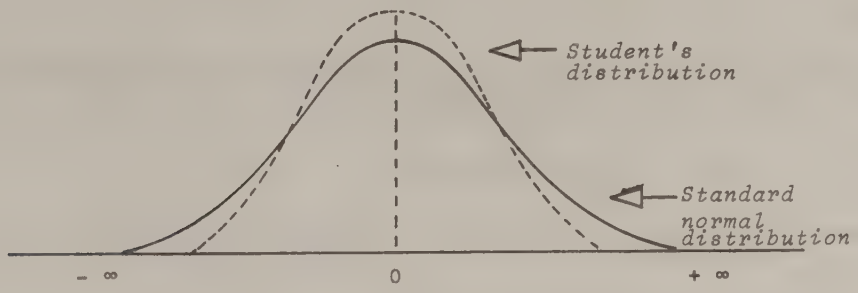


Figure 8.4.1

The tabulated values of the Student's distribution (see Appendix II) provides the values of t for given degrees of freedom (ν) and for specified levels of significance i.e.,

$$P(|t|>t_{\nu.a}) = a$$

or,

$$P(-t_{\nu.a}<t<+t_{\nu.a}) = 1-a$$

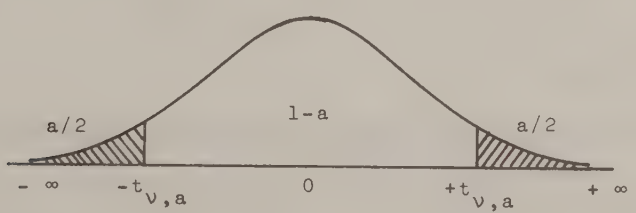


Figure 8.4.2

8.5 Confidence limits for a mean

Let x be normally distributed with mean m and variance σ^2. Let $\bar{x}$ and s^2 be their sample estimates based on a random sample of size n. Then, as section 8.4,

$$u = \frac{(\bar{x}-m)\sqrt{n}}{\sigma}$$

and,

$$v^2 = \frac{(n-1)s^2}{\sigma^2}$$

the ratio,

$$\frac{u\sqrt{n-1}}{v} = \{\frac{(\bar{x}-m)\sqrt{n}}{\sigma}\}\{\frac{\sigma}{s\sqrt{n-1}}\sqrt{n-1}\} = \frac{(\bar{x}-m)\sqrt{n}}{s} = \frac{(\bar{x}-m)}{s/\sqrt{n}}$$

possess a t distribution with $\nu = n-1$ degrees of freedom. If $t_{.05,\nu}$ represents the value of t such that the probability is .05 that $|t|>t_{.05,\nu}$, then the 95 percent confidence interval for m is given by,

$$\bar{x}-t_{.05,\nu}\frac{s}{\sqrt{n}}<m<\bar{x}+t_{.05,\nu}\frac{s}{\sqrt{n}}$$

Example

From a given pond containing fish of species A, a random sample of n = 10 fish was selected. The survey variate was the weight per fish (variate x, gr). The sample data are:

$$\sum_{i=1}^{10} x_i = 19950, \qquad \sum_{i=1}^{10} x_i^2 = 39800270$$

Estimate the confidence interval (95 percent) for the population mean.

Solution:

$$\bar{x} = 1995$$

$$s^2 = \frac{1}{n-1}\{\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{9}\} = \frac{1}{9}\{39800270 - \frac{(19950)^2}{10}\} = 2.22 gr^2$$

$$s/\sqrt{n} = \frac{\sqrt{2.22}}{\sqrt{10}} = 0.47$$

$$t_{.05,9} = 2.262$$

$$t_{.05,9} \cdot \frac{s}{\sqrt{n}} = 1.063$$

$$1995 - 1.063 < m < 1995 + 1.063$$

$$1993.937 gr < m < 1996.063 gr$$

8.6 Significance of the difference between two sample means $(\bar{x}_1 - \bar{x}_2)$

Assuming that two large samples of sizes n_1, n_2 have been selected from two non-normal populations with parameters m_1, σ_1^2, and m_2, σ_2^2 respectively, if $\bar{x}_1$ and $\bar{x}_2$ are the means of the selected samples, these estimators follow the normal distribution,

$$\bar{x}_1 : N(m_1, \frac{\sigma_1^2}{n_1}), \quad \bar{x}_2 : N(m_2, \frac{\sigma_2^2}{n_2})$$

It has been proved that the difference between the sample means $(\bar{x}_1 - \bar{x}_2)$ is a random variate which follows the normal distribution with,

Mean: $E(\bar{x}_1 - \bar{x}_2) = m_1 - m_2$

Variance: $V(\bar{x}_1 - \bar{x}_2) = \sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Standard error: $\sigma_{\bar{x}_1 - \bar{x}_2} = +\sqrt{\sigma^2_{\bar{x}_1 - \bar{x}_2}}$

The ratio, $t = \frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$

follows the standard normal distribution, t : N(0, 1). The 1-a confidence interval for t is:

$$-t_a < \frac{(\bar{x}_1 - \bar{x}_2) - (m_1 - m_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} < +t_a$$

From this inequality the confidence interval of (m_1-m_2), for a given probability level, can be established,

$$(\bar{x}_1-\bar{x}_2)-t_a\sigma_{\bar{x}_1-\bar{x}_2}<m_1-m_2<(\bar{x}_1-\bar{x}_2)+t_a\sigma_{\bar{x}_1-\bar{x}_2}$$

The theory presented above can also be used to test the hypothesis,

$$H_o : (m_1-m_2) = \delta_o, \quad (\delta_o \neq 0)$$

We accept the hypothesis if,

$$(\bar{x}_1-\bar{x}_2)-t_a\sigma_{\bar{x}_1-\bar{x}_2}<\delta_o<(\bar{x}_1-\bar{x}_2)+t_a\sigma_{\bar{x}_1-\bar{x}_2}$$

or if,

$$\frac{|(\bar{x}_1-\bar{x}_2)-\delta_o|}{\sigma_{\bar{x}_1-\bar{x}_2}} < t_a$$

that is, if the estimated t value is smaller than the corresponding significant t value (t_a).

Another field of application of the theory is to test the hypothesis,

$$H_o : (m_1-m_2) = \delta_o, \quad (\delta_o = 0)$$

or,

$$H_o : m_1 = m_2$$

In such a case the estimated t value is given by,

$$t = \frac{|\bar{x}_1-\bar{x}_2|}{\sigma_{\bar{x}_1-\bar{x}_2}}$$

The hypothesis $H_o : m_1 = m_2$ is valid, if the estimated t value is smaller than the corresponding significant t value (t_a). If $t>t_a$ the hypothesis is rejected. This means that the test has indicated a significant difference between the sample means, $\bar{x}_1-\bar{x}_2$.

8.7 Significance of the difference (m_1-m_2) when the population's variances are not known

If the population's variances are not known we cannot calculate the value of the sampling error of $\bar{x}_1-\bar{x}_2$, i.e.

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

An estimate of $\sigma_{\bar{x}_1-\bar{x}_2}$ can be calculated by using the sample data:

$$s_1^2 = \hat{\sigma}_1^2 = \frac{1}{n_1-1}\sum_i(x_{1i}-\bar{x}_1)^2$$

$$s_2^2 = \hat{\sigma}_2^2 = \frac{1}{n_2-1}\sum_i(x_{2i}-\bar{x}_2)^2$$

and,

$$s_{\bar{x}_1-\bar{x}_2} = \hat{\sigma}_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

a: If n_1 and n_2 are large (n_1, $n_2 > 50$) then the ratio,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

follows the standard normal distribution. The confidence interval for (m_1-m_2) for 1-a, is given by,

$$(\bar{x}_1-\bar{x}_2)-t_a s_{\bar{x}_1-\bar{x}_2} < m_1-m_2 < (\bar{x}_1-\bar{x}_2)+t_a s_{\bar{x}_1-\bar{x}_2}$$

Also, the estimated confidence interval can be used in order to test the hypothesis $H_o : m_1-m_2 = \delta_o$ (see previous section).

b: If the selected samples are small in size e.g. $n_1 = 10$, $n_2 = 15$, for testing the hypothesis $H_o : m_1-m_2 = \delta_o$, certain assumptions must be introduced in the experimental conditions of the test, i.e.

1. The two samples have been selected from normal populations.
2. The unknown variances of the populations are equal, $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

In such a case, if s^{*2} is the estimated pooled variance,

$$s^{*2} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{\sum_i (x_{1i}-\bar{x}_1)^2+\sum_i (x_{2i}-\bar{x}_2)^2}{n_1+n_2-2} ,$$

it can be proved that the ratio,

$$t = \frac{(\bar{x}_1-\bar{x}_2)-(m_1-m_2)}{s^* \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

follows the Student's t distribution ($\nu = n_1+n_2-2$). The estimated confidence interval for m_1-m_2 is,

$$(\bar{x}_1-\bar{x}_2)-t_{\nu,a} s^* \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < m_1-m_2 < (\bar{x}_1-\bar{x}_2)+t_{\nu,a} s^* \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Example

Juvenile fish of a given species were selected from two limnological zones of a lake. The table below gives the estimated weight per fish (x:gr) in the selected samples. Test the hypothesis that the average weight per fish in the two limnological zones is the same, $H_o : m_1 = m_2$.

- sample data -

Limnological zone	n_i	Weight per fish (gr)	$\sum x_i$ (gr)	$\sum x_i^2$ (gr^2)
L:1	$n_1=8$	8,10,10,9,11,9,7,7	$\sum x_{1i}=71$	$\sum x_{1i}^2=645$
L:2	$n_2=6$	6,10,7,9,12,8	$\sum x_{2i}=52$	$\sum x_{2i}^2=474$

Solution:

$$(n_1-1)s_1^2 = \{\sum_i x_{1i}^2 - \frac{(\sum_i x_{1i})^2}{n_i}\} = \{645-\frac{(71)^2}{8}\} = 14.875gr^2$$

$$(n_2-1)s_2^2 = \{\sum_i x_{2i}^2 - \frac{(\sum_i x_{2i})^2}{n_2}\} = \{474-\frac{(52)^2}{6}\} = 23.333gr^2$$

$$\overset{*}{s}^2 = \frac{14.875+23.333}{8+6-2} = 3.184gr^2$$

$$\overset{*}{s} = +\sqrt{3.184gr^2} = 1.7844gr$$

$$s_{\bar{x}_1-\bar{x}_2} = 1.7844\sqrt{\frac{1}{8}+\frac{1}{6}} = 0.9636gr$$

Estimated Student's t value,

$$t = \frac{|\bar{x}_1-\bar{x}_2|}{s_{\bar{x}_1-\bar{x}_2}} = \frac{|8.8750-8.6667|}{0.9636} = 0.2159$$

Tabulated Student's t value,

$$t_{\nu=12,\ a=0.05} = 2.179$$

Since,

$$t<t_{\nu.a},\quad 0.2159<2.179$$

the established hypothesis, $H_o : m_1 = m_2$ is valid.

8.8 Non-independent samples (t-student)

When inherent pairing exists between the pairs of items in two samples, it usually follows that the two samples are not independent. It is extremely important that non-independent samples be recognized as such; they must not be treated as independent samples.

Example

In order to test the effect of the trawling period on fish catch a pair of trawlers was used, A and B. The trawling period of A was half an hour and of B one hour. The obtained sample data converted into kg/haul/hour are given below. Test the significance of the difference between the average catches per haul for A and B.

-Catches per haul/hour (kg)-

Trawler A	y_i:	300	290	450	360	270	420	330
Trawler B	x_i:	320	304	370	420	340	390	340
D = A-B	$d_i=y_i-x_i$	-20	-14	+80	-60	-70	+30	-10

In order to test the above hypothesis we determine the value of $\bar{d}$ and ascertain whether $\bar{d}$ differs significantly from zero. Here, the null hypothesis is that $\bar{d}$ is the mean of a random sample from a population of differences having a mean of zero:

a) $d_i = y_i-x_i$

b) $\bar{d} = \frac{1}{n}\sum d_i = -9.1429$

c) $s_d^2 = \frac{1}{n-1}\{\sum d_i^2 - \frac{(\sum d_i)^2}{n}\} =$

$$\frac{1}{6}\{16496 - \frac{(-64)^2}{7}\} = 2651.81$$

d) $s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{51.50}{2.65} = 19.43$

e) $t = \frac{|\bar{d}|\sqrt{n}}{s_d} = \frac{|9.1429| \times 2.65}{51.50} = 0.47$

f) $t_{\nu=6,\ a=0.05} = 2.447$

g) Conclusion: The estimated t value is smaller than the tabulated value, the established hypothesis is valid, i.e. there are no significant differences between the average catches per haul between A and B.

8.9 Distribution of proportion $p = \frac{x}{n}$ (large samples)

Assuming that, in order to estimate the proportion of fish in a fish pond which suffer from a given disease (P), we select a series of samples of size n, the variate x expresses the number of fish in the samples suffering from the disease (binomial variate). It has been proved that,

$$E(x) = nP$$
$$V(x) = nP(1-P)$$

If p is the sample value of the proportion ($p = \hat{P}$), p follows the binomial distribution,

$$E(p) = \frac{1}{n}E(x) = \frac{1}{n} \times nP = P$$

$$V(p) = \sigma_p^2 = \frac{1}{n^2}V(x) = \frac{1}{n^2}\, nP(1-P) = \frac{P(1-P)}{n}$$

It has been proved that as $n \to \infty$ the ratio,

$$t = \frac{x - E(x)}{\sigma_x} = \frac{x - nP}{\sqrt{nP(1-P)}} = \frac{\frac{x}{n} - P}{\sqrt{\frac{P(1-P)}{n}}} = \frac{p-P}{\sigma_p}$$

tends to the standard normal variate. Specifically, for large samples e.g. n>200, the sample proportion p follows the normal distribution,

$$p : N(P, \sigma_p = \sqrt{\frac{P(1-P)}{n}})$$

It can be proved that the estimated confidence interval of P (for a given probability) is given by,

$$p - t_a s_p < P < p + t_a s_p$$

where,

$$s_p = \sqrt{\frac{p(1-p)}{n}}$$

Also, the estimated confidence interval can be used to test the hypothesis, H_o : $P = P_o$.

Example

In order to test the hypothesis of an even distribution of males and females in a fishing area, a sample of 400 fish was selected. The sample proportion of male fish in the sample was p = 0.44. Test the hypothesis that H_o : P = 0.50, (a = 0.05),

Solution:

$$p = 0.44, \quad s_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.44\times 0.56}{400}} = 0.0248$$

Estimated t value,

$$t = \frac{|0.44-0.50|}{0.0248} = 2.42$$

Tabulated $t_{a=0.05} = 1.96$

Since $t > t_a$, $2.42 > 1.96$ the hypothesis (H_o : P = 0.50) is discredited.

The confidence interval for P is given by,

$$0.44-1.96\times 0.0248 < P < 0.44+1.96\times 0.0248$$

$$0.3914 < P < 0.4886$$

8.10 Distribution of the difference (p_1-p_2)

If from a given population we select two large samples of sizes n_1, n_2 and in each selected sample we calculate the proportions,

$$p_1 = \frac{x_1}{n_1}, \quad p_2 = \frac{x_2}{n_2}$$

then the difference (p_1-p_2) follows the normal distribution with,

$$\text{Mean, } m_{p_1-p_2} = E(p_1-p_2) = 0$$

$$\text{Variance, } \sigma^2_{p_1-p_2} = P(1-P)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

An estimate of P is given by,

$$\overset{*}{p} = \hat{P} = \frac{n_1p_1+n_2p_2}{n_1+n_2} = \frac{x_1+x_2}{n_1+n_2}$$

An estimate of $\sigma^2_{p_1-p_2}$ is given by,

$$s^2_{p_1-p_2} = \overset{*}{p}(1-\overset{*}{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

The confidence interval for p_1-p_2 (= 0) is given by (for given 1-a),

$$(p_1-p_2)-t_a s_{p_1-p_2} < P_1-P_2 < (p_1-p_2)+t_a s_{p_1-p_2}$$

Also, the established confidence interval can be used to test the hypothesis,

H_o : $P_1 = P_2$

i.e. The two samples have been selected from the same population.

Example

In fish farming studies two different kinds of treatment were used to protect the fish from a given disease. Treatment-1 was applied on a sample of $n_1 = 120$ fish and treatment-2 was applied on a sample of $n_2 = 160$ fish. The estimated proportions of fish affected by the disease in the two samples are respectively,

$$p_1 = 0.082, \quad p_2 = 0.094$$

Test the hypothesis that $H_o : P_1 = P_2$, $(a = 0.05)$

Solution:

$$p^* = \frac{120\times0.082+160\times0.094}{280} = 0.0889$$

$$s^2_{p_1-p_2} = 0.0889\times0.9111\left(\frac{1}{120}+\frac{1}{160}\right) = 0.0012$$

Estimated t value,

$$t = \frac{|0.082-0.094|}{\sqrt{0.0012}} = 0.346$$

Tabulated t value,

$$t_{a=0.05} = 1.96$$

Since $t<t_{a=0.05}$, $0.346<1.96$, the hypothesis is valid. This means that the two treatments do not produce significantly different results.

9. THE χ^2 TEST

9.1 The χ^2 variate

Assuming that from a normal population $N(m, \sigma^2)$ we select random samples of size n, $x_1, x_2, \ldots, x_n$, and that in each selected sample we calculate the value of the following function:

$$\left(\frac{x_1-m}{\sigma}\right)^2+\left(\frac{x_2-m}{\sigma}\right)^2+ \ldots +\left(\frac{x_n-m}{\sigma}\right)^2$$

The above function defines a random variable, χ^2. The values of χ^2 variate range from 0 to ∞.

$$\chi^2 = \sum_{i=1}^{n}\left(\frac{x_i-m}{\sigma}\right)^2, \quad 0\leq\chi^2\leq\infty$$

In the above function the number of independent terms denote the degrees of freedom of the χ^2 variate, d.f. $\nu = n$. The distribution of the χ^2 variate depends on the degrees of freedom. The figure below (Figure 9.1.1) gives the distribution of χ^2 for various degrees of freedom[1/].

1/ The mode of the χ^2 distribution is at $\chi^2 = \nu-2$, except when $n = 1$, and then the mode is at zero; the mean is at $\chi^2 = \nu$ and the variance 2ν.

When $\nu>30$, Fisher has shown that $\sqrt{2\chi^2}$ is approximately normally distributed about mean $\sqrt{2\nu-1}$ with unit variance. Thus $\sqrt{2\chi^2} - \sqrt{2\nu-1}$ may be considered as a standardized normal variate for values $\nu>30$.

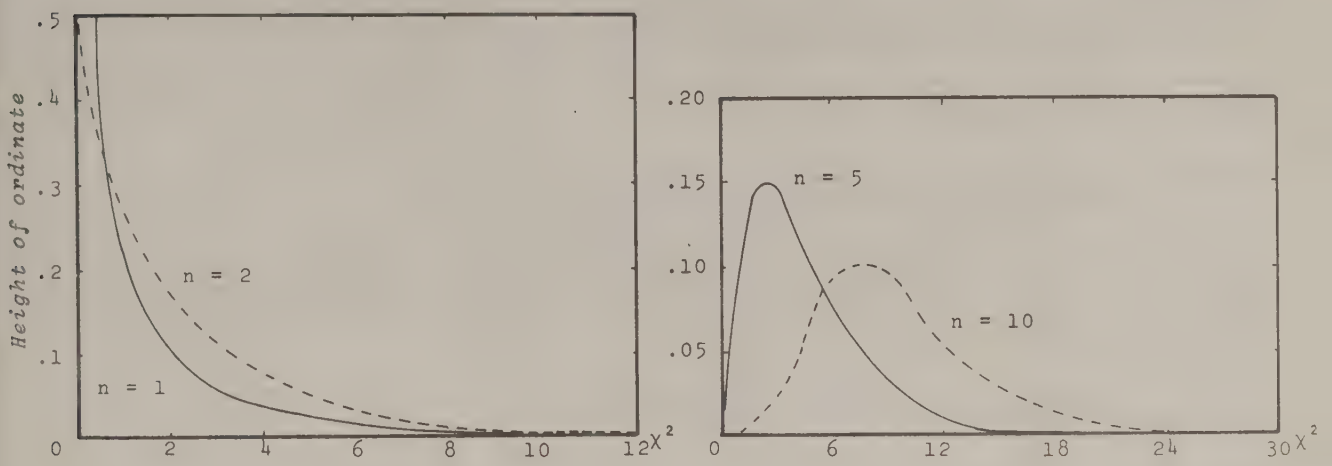


Figure 9.1.1 The χ^2 distribution for n = 1, n = 2, n = 5, n = 10

In Appendix II the table of χ^2 distribution gives the critical values of χ^2_α which the χ^2 exceeds with a given probability,

$$P(\chi^2 \geq \chi^2_\alpha) = a$$

The respective χ^2_α values correspond in a : 0.01, 0.02, 0.05, 0.10, 0.90, 0.95, 0.99 (see also Figure 9.1.2).

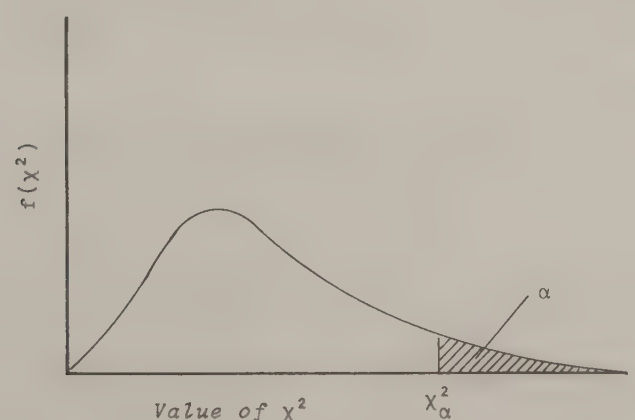


Figure 9.1.2

Additive property: It has been proved that if u_1 and u_2, independent random variables, are distributed like χ^2 with ν_1 and ν_2 degrees of freedom respectively, then, $u = u_1+u_2$ is distributed like χ^2 with $\nu = \nu_1+\nu_2$ d.f.

9.2 The estimator of the variance σ^2

If $\bar{x}$ is the sample mean of the variate X, then, the function,

$$\frac{1}{n}\{(x_1-\bar{x})^2+(x_2-\bar{x})^2+ \ldots +(x_n-\bar{x})^2\}$$

is a function of the sample values and is taken as an estimator of the population variance σ^2. Specifically,

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Variances of random samples from a normal distribution are distributed neither normally nor symmetrically. Their distribution follows a skewed curve (skewed to the right), the exact shape of which depends upon σ^2 and n. It has been proved, however, that the function,

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n}(x_i-\bar{x})^2/\sigma^2 = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2$$

follows the χ^2 distribution with $\nu = n-1$ d.f.

9.3 Significance of the difference between s^2 and σ^2

In the littoral zone of a lake the variance of the length variate of species A is $\sigma^2 = 46\text{cm}^2$. A random sample of n = 20 fish (species A) was selected from the lacustrine zone of the lake. The calculated sample variance of the survey variate was $s^2 = 54\text{cm}^2$. Test the hypothesis that $s^2 = 54\text{cm}^2$ is the variance of a random sample from a population having $\sigma^2 = 46\text{cm}^2$.

Solution:

Computing χ^2, we find:

$$\chi^2 = \frac{(20-1)54}{46} = 22.3043$$

From the table of χ^2 distribution it is seen that if $\sigma^2 = 46\text{cm}^2$ the probability of obtaining $s^2 = 54\text{cm}^2$ or larger, for samples of n = 20, is almost exactly 30 percent (one tail test). The hypothesis is not discredited ($\chi^2_{.05,19} = 30.144$).

9.4 Confidence limits for σ^2

We may also employ χ^2 to obtain the confidence limits of σ^2. By using deck sampling (n = 20) the variance of the length variate of fish of a given species was estimated, $s^2 = 75\text{cm}^2$. What are the 90 percent confidence limits of σ^2?

Solution:

To estimate the confidence limits for σ^2, we use two χ^2 values from the table of χ^2 distribution for $\nu = 20-1 = 19$ d.f. One at the upper 0.05 point and one at the lower 0.05 point (the 0.95 in the table). These χ^2 values are, respectively,

$$\chi^2_1 = 30.144, \quad \chi^2_2 = 10.117,$$

and we solve,

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \text{ for } \sigma^2$$

$$30.144 = \frac{(19)(75)}{\sigma^2_1}$$

$$10.117 = \frac{(19)(75)}{\sigma^2_2}$$

The 90 percent confidence limits for σ^2 are 47.2731 and 140.8520 (the corresponding confidence limits for the population standard deviation ($\sigma = +\sqrt{\sigma^2}$) are, 6.8755 and 11.8681). The estimated confidence limits for σ^2 indicate that if we compute many such 90 percent limits from random samples from a normal population, our statements will include the population value 90 percent of the time and fail to include it 10 percent of the time.

9.5 Testing goodness of fit

A problem that arises frequently in statistical work is the testing of compatability of a set of observed and theoretical frequencies. Here, the hypothesis to be tested can be explained in simple language as follows. Assume that a random sample of size n has been selected and that the frequency distribution of the sample data is,

$$\begin{vmatrix} x_0-x_1 & f_1 \\ x_1-x_2 & f_2 \\ x_2-x_3 & f_3 \\ & \\ & \\ x_{k-1}-x_k & f_k \end{vmatrix} \qquad n = \sum_i f_i$$

By using the sample frequency distribution, test the hypothesis that the sample has been selected from a population of which the frequency function f(x) is described by the mathematical model $f_0(x)$.

To test the hypothesis the following procedure is employed:

1. By using the sample observations estimates are calculated for the parameters of $f_0(x)$.
2. By using the estimated $f_0(x)$ we calculate the probabilities of the established class intervals of the variate (p_i) and the theoretical frequencies $\theta_i = np_i$.
3. We calculate the value of,
$$\chi^2 = \sum \frac{(f_i-\theta_i)^2}{\theta_i}$$
4. The estimated χ^2 is compared[1/] with the tabulated $\chi^2_{\nu,\alpha}$. The hypothesis is valid if,
$$\chi^2 < \chi^2_{\nu,\alpha}$$
The hypothesis is discredited if,
$$\chi^2 > \chi^2_{\nu,\alpha}$$

Example

The following table gives the empirical length frequency distribution of n = 500 fish and the corresponding theoretical frequencies θ_i (normal variate). Test the hypothesis that the sample was a random one from a normal population.

1/ It should be noted that, since the χ^2 curve is an approximation to the discrete χ^2 frequency function care must be exercised that the χ^2 test is used only when the approximation is good. Experience and theoretical investigations indicate that the approximation is usually satisfactory, provided that the frequencies of the class intervals are ≥ 5 and that the number of classes in the frequency distribution are ≥ 5.

Table 9.5.1 χ^2 test of goodness of fit for normal curve to length fish distribution

Length x (cm)	f_i	θ_i	$f_i-\theta_i$	$\frac{(f_i-\theta_i)^2}{\theta_i}$	Remarks
-5	4	6			
5-10	40)44	47)53	-9	1.528	) Combined
10-15	150	148	2	0.027	
15-20	200	186	14	1.054	
20-25	85	93	-8	0.688	
25-30	21	20	1	0.050	
	500	500		$\chi^2=3.347$	

In Table 9.5.1 the number of classes of the distribution is k = 5. The degrees of freedom for the tabulated chi-square are $\nu = 5-1-2$. The tabulated value of chi-square is,

$$\chi^2_{\nu=2;\ \alpha=0.05} = 5.991$$

Since $\chi^2<\chi^2_{\nu,\alpha}$, $3.347<5.991$ the hypothesis is valid.

It should be noted that in the above example the parameters m, σ^2 were estimated by using the sample observations. In such a case the calculated χ^2 is distributed with $\nu = k-1-\S$ degrees of freedom, where $\S=2$ expresses the number of the estimated parameters.

The same technique as above is used when we test the hypothesis that the sample has been selected from a population which follows the Poisson ($\S=1$) or binomial distribution ($\S=2$).

10. DESCRIBING A FREQUENCY DISTRIBUTION BY A FITTED MATHEMATICAL MODEL

A problem that arises frequently in statistical work is the fitting of a curve to a frequency distribution. The purpose of fitting a curve to a frequency distribution may be either to ascertain whether a given curve describes the general shape of a distribution or to fit a curve to values obtained from repeated samples taken from the same population.

10.1 Fitting a normal curve

If one has reasons for believing that a set of data represents a random sample from some normal population, then the fitted normal curve would serve as an approximation to the population curve. For example, it has been observed that the length distribution of trout fish of the same year class follows the normal distribution. In such a case a large sample of fish from the survey population would reveal the shape of the distribution. Since a normal distribution is completely determined by its mean and standard deviation and these quantities can be estimated from the sample data, one would have considerably more confidence in the fitted normal curve as representing the population distribution rather than the histogram of the empirical frequency distribution.

An illustration of the technique of fitting a normal curve (areas) to an empirical frequency distribution is given in the following table (Table 10.1.1). The empirical data represent the length distribution of a sample of 520 trout of a given year class.

Table 10.1.1 Length distribution of trout of a given year class

Length (x) - cm - (1)	No.of fish (sample data) (2)	$t_i=\frac{x_i-\bar{x}}{s}$ (3)	$P(t<t_i)$ (4)	$F_i=n\times(4)$ (5)	$\theta_i=F_i-F_{i-1}$ (6)	θ_i=adjusted theor. frequencies (7)
22-23	5	-2.32	0.01017	5.2884	5.2884	5
23-24	7	-1.96	0.02500	13.0000	7.7116	8
24-25	18	-1.60	0.05480	28.4960	15.4960	16
25-26	30	-1.24	0.10749	55.8948	27.3988	27
26-27	45	-0.87	0.19215	99.9180	44.0232	44
27-28	55	-0.52	0.30153	156.7956	56.8776	57
28-29	70	-0.16	0.43644	226.9488	70.1532	70
29-30	75	0.20	0.57926	301.2152	74.2637	74
30-31	65	0.56	0.71226	370.3752	69.1627	69
31-32	55	0.92	0.82121	427.0292	56.6540	57
32-33	40	1.28	0.89973	467.8596	40.8304	41
33-34	25	1.64	0.94950	493.7400	25.8804	26
34-35	12	2.00	0.97725	508.1700	14.4300	15
35-36	10	2.36	0.99086	515.2472	7.0772	7
36-37	8	2.72	0.99674	519.3048	4.0576	4
	n=520					n=520

In the above table (Table 10.1.1) the values of $\bar{x}$ and s were calculated by using the sample data:

$$\bar{x} = \frac{1}{n}\sum_i f_i x_i = 29.44\text{cm}, \quad s^2 = \frac{1}{n-1}\left(\sum_i f_i x_i^2 - \frac{\{\sum_i f_i x_i\}^2}{n}\right) = 7.71\text{cm}$$

and, $s = +\sqrt{s^2} = 2.78\text{cm}$.

The columns of the table express:

(3) The values of the normal standard variate ($x_i = x_{max.}$).

(4) The proportion of fish with length less than t_i, (cumulative area).

(5) The theoretical number of fish with length less than t_i (cumulative number).

(6) The theoretical number of fish by length class interval (unadjusted).

(7) The theoretical number of fish by length class interval (adjusted).

10.2 Fitting a binomial

It has been proved that there are certain criteria which can be used to determine the mathematical model describing an empirical frequency distribution (discrete variables). Specifically, if m and σ^2 are the mean and variance of the survey variate then:

1. A binomial model (positive) can be fitted to the empirical frequency distribution if $\sigma^2<m$.
2. A Poisson model can be fitted to the empirical frequency distribution if $\sigma^2 = m$.
3. A negative binomial can be fitted to the empirical frequency distribution if $\sigma^2>\mu$.

The following table (Table 10.2.1) gives the area distribution of species A in a given lake (hypothetical data), and illustrates the technique of fitting a binomial model to the empirical frequency distribution.

Table 10.2.1 Area distribution of species A in a given lake

Number of fish/area unit, x (1)	Number of area units, f_i (2)	3=1x2 (3)	P_x* (4)	Unadjusted theoretic. frequencies nP_x (5)	Adjusted theoretic. frequencies (6)
0	85	0	0.11765	58.825	59
1	100	100	0.30253	151.265	151
2	200	400	0.32414	162.070	162
3	80	240	0.18524	92.620	92
4	20	80	0.05941	29.705	30
5	10	50	0.01207	6.035	6
6	5	30	0.00086	0.430	0
	n=500	900			500

* See Chapter 7.

The estimated mean and variance of the empirical distribution are $\bar{x} = 1.8$ fish and $s^2 = 1.5$ fish^2 respectively, ($s^2 < \bar{x}$ or $1.5 < 1.8$).

The columns in Table 10.2.1 express:

(4) The estimated probabilities of the binomial model (P_x).

(5) Estimated unadjusted theoretical frequencies.

(6) Estimated adjusted theoretical frequencies.

10.3 Fitting a Poisson

The table below (Table 10.3.1) gives the area distribution of large fish in a lake (hypothetical data) and illustrates the technique of fitting a Poisson model to the empirical frequency distribution.

Table 10.3.1 Area distribution of large fish in a lake

No.of fish /area unit - x - (1)	No.of area units, f_i (2)	(3)=1x2 (3)	P_x* (4)	Unadjusted theoretic. frequencies nP_x (5)	Adjusted theoretic. frequencies (6)
0	43	0	0.074274	44.56	45
1	110	110	0.193112	115.87	116
2	200	400	0.251046	150.83	151
3	100	300	0.217573	130.54	130
4	70	280	0.141422	84.85	85
5	34	170	0.073540	44.10	44
6	19	114	0.031867	19.08	19
7	12	84	0.011813	7.08	7
8	7	56	0.004138	2.46	2
9	4	36	0.001195	0.72	1
10	1	10	0.000311	0.18	0
	n=600	1560			600

* See Chapter 7.

The columns of the table express:

(4) The estimated probabilities of the Poisson model.

(5) Estimated unadjusted theoretical frequencies.

(6) Estimated adjusted theoretical frequencies.

11. ANALYSIS OF VARIANCE[1/]

In designing an experiment, the experimenter usually has in mind the testing of a hypothesis or the estimation of some parameters. The analysis of variance enables the experimenter to design sensitive experiments for either of these basic problems. However, in our case the explanation of the technique is made from the point of view of testing hypothesis.

11.1 One criterion of classification

The problem: A fishery statistician wanted to know whether the mean catchability of bottom gillnets fished in three different depths of water, W_1, W_2, W_3 are significantly different from each other. For this purpose fifteen similar gillnet units were set. Specifically, five points were randomly selected within each W_i area ($i = 1, 2, 3$) and the respective gillnet units were fished by using identical experimental conditions. The obtained fish catch per gillnet unit is given below (x:lbs/day).

-fish catch/gillnet unit (lbs)/day-

	Area - W_1 ($i=1$)	Area - W_2 ($i=2$)	Area - W_3 ($i=3$)	Total	Remarks
$j=1$	15	16	13		
$j=2$	14	14	12		
$j=3$	12	13	11		
$j=4$	13	15	12		
$j=5$	13	14	10		
T_i	67	72	58	T:197	(sample totals)
$\bar{x}_{i.}$	13.4	14.4	11.6	$\bar{x}$:13.13	(sample means)

Symbols:

suffix i : stands for the areas (columns) ($i = 1, 2, \ldots, h$), ($h = 3$).

suffix j : stands for the values of the variate within the columns ($j = 1, 2, \ldots, n_i$)

$\bar{x}_i = \frac{1}{n_i}\sum_j x_{ij} = \frac{T_{i.}}{n_i}$: Mean within the i^{th} column ($T_{i.}$ = column total, $n_i = 5$)

$\bar{x} = \frac{1}{n}\sum_i\sum_j x_{ij} = \frac{T}{n}$: Overall mean (T = overall total, $n = \sum_i n_i = 15$)

Assumptions made in analysis of variance:

1. The samples are random.
2. The samples are independent.
3. The populations are normal.
4. The populations all have the same variance (often called the assumption of homoscedasticity).

Total variation = variation between columns + variation within columns.

Assuming that the classifying factor (W) does not influence variate values, each column into which the sample is divided by this factor will be a random sample from the parent population. If σ^2 is the variance of the parent population, then, two unbiased estimates of σ^2 are given by,

---

1/ Components of variance model.

$$s_b^2 = \frac{1}{h-1}\sum_{i=1}^{h} n_i(\bar{x}_{i.}-\bar{\bar{x}})^2$$

and,

$$s_w^2 = \frac{1}{n-h}\sum_{ij}(x_{ij}-\bar{x}_{i.})^2$$

specifically,

$$\sum_{ij}(x_{ij}-\bar{\bar{x}})^2 = \sum_{ij}(x_{ij}-\bar{x}_{i.}+\bar{x}_{i.}-\bar{\bar{x}})^2$$

$$= \sum_{ij}(x_{ij}-\bar{x}_{i.})^2+\sum_{ij}(\bar{x}_{i.}-\bar{\bar{x}})^2+2\times 0$$

$$= \sum_{ij}(x_{ij}-\bar{x}_{i.})^2+\sum_{i} n_i(\bar{x}_{i.}-\bar{\bar{x}})^2$$

and,

$$(n-1)s^2 = (n-h)s_w^2+(h-1)s_b^2$$

degrees of freedom,

$$n-1 = (n-h)+(h-1)$$

Now we impose the restriction that the population sampled be normal. With this restriction, the two unbiased estimates are also independent, and all the conditions for applying the F-test[1/] are fulfilled.

Algebraical transformations:

a. $$\sum_{ij}(x_{ij}-\bar{\bar{x}})^2 = \sum_{ij}x_{ij}^2-\frac{(\sum_{ij}x_{ij})^2}{n}, \text{ if } T = \sum_{ij}x_{ij}$$

then,

$$\sum_{ij}(x_{ij}-\bar{\bar{x}})^2 = \sum_{ij}x_{ij}^2-\frac{T^2}{n}$$

b. $$\sum_{i} n_i(\bar{x}_{i.}-\bar{\bar{x}})^2 = \sum_{i} n_i(\bar{x}_{i.}^2+\bar{\bar{x}}^2-2\bar{\bar{x}}\bar{x}_{i.})$$

$$= \sum_{i} n_i\bar{x}_{i.}^2+n\bar{\bar{x}}^2-2\bar{\bar{x}}\sum_{i} n_i\bar{x}_{i.}$$

If $T_{i.} = \sum_{j} x_{ij}$ then $\bar{x}_{i.} = \frac{T_{i.}}{n_i}$

also, $\bar{\bar{x}} = \frac{T}{n}$ and $\sum_{i} n_i\bar{x}_{i.} = T$

1/ When s_1^2 and s_2^2 are independent estimates of σ^2 from the same normal population based on samples n_1 and n_2 respectively, their ratio,

$$\frac{s_1^2}{s_2^2}$$

is distributed according to F distribution with $\nu_1 = n_1-1$ and $\nu_2 = n_2-1$ degrees of freedom.

If $s_1^2 = s_2^2$ the value of F = 1.0. Values of F vary from 0 to 0.999... when $s_1^2<s_2^2$ and from 1.000... 1 to ∞ when $s_1^2>s_2^2$. Values of F for selected values of ν_1 and ν_2 and for probabilities 0.05 and 0.01 in the right tail of the distribution are given in Appendix II.

$$\sum_i n_i(\bar{x}_{i\cdot}-\bar{\bar{x}})^2 = \sum_i n_i \frac{T_{i\cdot}^2}{n_i^2} + n\frac{T^2}{n^2} - 2\frac{T}{n}\cdot T$$

and,

$$\sum_i n_i(\bar{x}_{i\cdot}-\bar{\bar{x}})^2 = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{n}$$

c. $$\sum_{ij}(x_{ij}-\bar{x}_{i\cdot})^2 = \sum_i \{\sum_j x_{ij}^2 - \frac{(\sum_j x_{ij})^2}{n_i}\}$$

and,

$$\sum_{ij}(x_{ij}-\bar{x}_{i\cdot})^2 = \sum_{ij} x_{ij}^2 - \sum_i \frac{T_{i\cdot}^2}{n_i}$$

Table of analysis of variance for one criterion of classification

Source of variation	Degrees of freedom	Sum of squares		Estimate of variance
1.Between columns	h-1	$\sum_i n_i(\bar{x}_{i\cdot}-\bar{\bar{x}})^2$	or $\sum_i \frac{T_{i\cdot}^2}{n_i} - \frac{T^2}{n}$	$s_b^2 = \frac{1}{h-1}(\sum_i \frac{T_{i\cdot}^2}{n_i} - \frac{T^2}{n})$
2.Within columns	n-h	$\sum_{ij}(x_{ij}-\bar{x}_{i\cdot})^2$	or $\sum_{ij} x_{ij}^2 - \sum_i \frac{T_{i\cdot}^2}{n_i}$	$s_w^2 = \frac{1}{n-h}(\sum_{ij} x_{ij}^2 - \sum_i \frac{T_{i\cdot}^2}{n_i})$
3.Total	n-1	$\sum_{ij}(x_{ij}-\bar{\bar{x}})^2$	or $\sum_{ij} x_{ij}^2 - \frac{T^2}{n}$	———

In our example the table of analysis of variance is,

Table of analysis of variance for one criterion of classification

Source of variation	d.f.	Sum of squares		Estimate of variance
1.Between columns	h-1=2	$\frac{1}{5}\sum_i T_{i\cdot}^2 - \frac{T^2}{n}$	20.13 = 2607.40-2587.27	$s_b^2 = 10.06\ \text{lbs}^2$
2.Within columns	n-h=12	$\sum_{ij} x_{ij}^2 - \frac{1}{5}\sum_i T_{i\cdot}^2$	15.60 = 2623-2607.40	$s_w^2 = 1.30\ \text{lbs}^2$
3.Total	n-1=14	$\sum_{ij} x_{ij}^2 - \frac{T^2}{n}$	35.73 = 2623-2587.27	———

The F-test is the ratio,

$$F = \frac{s_b^2}{s_w^2} = \frac{10.06\text{lbs}^2}{1.30\text{lbs}^2} = 7.733$$

The tabulated F for significance ($\nu_1 = 2$ and $\nu_2 = 12$) at the 5 percent level is 3.88, and that for 1 percent level is 6.93. As the calculated value of F exceeds 6.93 we conclude that there is a highly significant difference in the catchability level of gillnet units between fishing grounds of different depths (it would be more precise to say that we reject the hypothesis that there is no difference in the catchability level of the gillnet units between fishing grounds of different depths)1/.

What is different from what? Having shown that there is a highly significant difference in the level of catchability of the gillnet units between fishing grounds of different depths, it is logical to ask if this is due entirely to the comparatively low catches in W_3, or are catches in W_1 and W_2 also different? The averages $\bar{x}_{1.}$, $\bar{x}_{2.}$, $\bar{x}_{3.}$ may be compared by means of a t-test.

To make the t-test, we have to estimate the standard error of the mean and to get that we need an estimate of s^2. This we already have: it is the s_w^2.

(1) Testing $\bar{x}_{1.}$ vs $\bar{x}_{2.}$

$$t = \frac{|\bar{x}_{1.} - \bar{x}_{2.}|}{s_{\bar{x}_{1.} - \bar{x}_{2.}}}$$

where,

$$s_{\bar{x}_{1.} - \bar{x}_{2.}} = s_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{1.30}\sqrt{\frac{2}{5}} = 0.72$$

Substituting we get,

$$t = \frac{|13.40 - 14.40|}{.72} = 1.389 \text{ - not significant}$$

(2) Testing $\bar{x}_{1.}$ vs $\bar{x}_{3.}$

$$t = \frac{|13.40 - 11.60|}{.72} = 2.5 \text{ - significant at 5 percent level}$$

(3) Testing $\bar{x}_{2.}$ vs $\bar{x}_{3.}$

$$t = \frac{|14.40 - 11.60|}{.72} = 3.889 \text{ - significant at 1 percent level}$$

Note: The tabulated values of t for $\nu = 12$ d.f. are, at 1 percent level 3.055, at 5 percent level 2.179.

11.2 Two criteria of classification (one individual in each sub-class)

Now let us consider the case where there are two criteria of classification. Suppose a lake has been divided into three "fishing areas" and that four tribes, A, B, C, D, are fishing on the lake. To examine whether fish catch varies significantly with fishing area and tribe, we must analyze the variation in fish catch into three components, that due to the fishing area, that due to variety of tribes, and, finally, the residual variations (errors) due to unspecified or unknown causes assumed to be normal.

1/ If F had been smaller than 6.93 but larger than 3.88, we would have said that the difference is significant (instead of "highly significant"). If F had been less than 3.88 we would have said that the difference in the catchability level of the gillnet units between the fishing grounds is not significant.

Supplementary notation:

k : number of rows (j = 1, 2, ... k)

$T_{.j}$: sample total of a given row

$\bar{x}_{.j} = \frac{T_{.j}}{n_j}$, row mean ($n_j$: number of sample observations in the j^{th} row).

Table of analysis of variance for two criteria of classification

Source of variation	d.f.	Sum of squares		Estimate of variance
1.Between columns	h-1	$A=k\sum_i(\bar{x}_{i.}-\bar{x})^2$	$\sum_i \frac{T_{i.}^2}{n_i}-\frac{T^2}{n}$	$s_3^2=\frac{1}{h-1}(\sum_i \frac{T_{i.}^2}{n_i}-\frac{T^2}{n})$
2.Between rows	k-1	$B=h\sum_j(\bar{x}_{.j}-\bar{x})^2$	$\sum_j \frac{T_{.j}^2}{n_j}-\frac{T^2}{n}$	$s_2^2=\frac{1}{k-1}(\sum_j \frac{T_{.j}^2}{n_j}-\frac{T^2}{n})$
3.Error	(h-1)(k-1)	$C=\sum_i\sum_j d_{ij}^2$ 1/	1/	$s_1^2=\frac{1}{(h-1)(k-1)}\sum_i\sum_j d_{ij}^2$
4.Total	n-1	$D=\sum_i\sum_j(x_{ij}-\bar{x})^2$	$\sum_i\sum_j x_{ij}^2-\frac{T^2}{n}$	——

1/ $\sum_i\sum_j d_{ij}^2 = \sum_i\sum_j(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{x})^2$

$$= \sum_i\sum_j x_{ij}^2-\frac{T^2}{n}-\{(\sum_i \frac{T_{i.}^2}{n_i}-\frac{T^2}{n})+(\sum_j \frac{T_{.j}^2}{n_j}-\frac{T^2}{n})\}$$

or,

$$\sum_i\sum_j d_{ij}^2 = D-(A+B)$$

In the following table the data of the experiment described above are given,

- kg/FEU/day -

Tribes	Fishing areas I (i=1)	II (i=2)	III (i=3)	$T_{.j}$	$\frac{T_{.j}^2}{n_j}$	$\bar{x}_{.j}$	Remarks
j=1 A	20	24	30	74	1825.33	24.67	$\sum_i\sum_j x_{ij}^2$=11018
j=2 B	22	27	34	83	2296.33	27.67	
j=3 C	23	29	37	89	2640.33	29.67	$\frac{T^2}{n}$ = 10443
j=4 D	28	35	45	108	3888.00	36.00	
$T_{i.}$:	93	115	146	(T:354)	10649.99		
$\frac{T_{i.}^2}{n_i}$:	2162.25	3306.25	5329	10797.50			
$\bar{x}_{i.}$:	23.25	28.75	36.50				

Substituting the calculated values in the table of analysis of variance for two criteria of classification we get,

Source of variation	d.f.	Sum of squares	Estimate of variance
1.Between columns	2	354.50 = 10797.50-10443	$s_3^2 = \frac{354.50}{2} = 177.25$
2.Between rows	3	206.99 = 10649.99-10443	$s_2^2 = \frac{206.99}{3} = 68.99$
3.Error	6	13.51 = 575-(354.50+206.99)	$s_1^2 = \frac{13.51}{6} = 2.25$
4.Total	11	575 = 11018-10443	———

Estimated F values:

1. Between columns: $F_1 = \frac{177.25}{2.25} = 78.78$

2. Between rows: $F_2 = \frac{68.99}{2.25} = 30.66$

Tabulated F values (5 percent, 1 percent):

1. $F_{2,6} = 5.143, 10.925, (F_1 > F_{2.6})$
2. $F_{3,6} = 4.757, 9.779, (F_2 > F_{3.6})$

Conclusions:

1. The difference in fish catch (kg/FEU/day) is highly significant between the fishing areas.
2. The difference in fish catch is highly significant between the tribes.

What is different from what? Having shown that there are highly significant differences in fish catch between fishing areas as well as between tribes it is now logical to test the significance between the estimated means.

1. Testing the significance of the means between columns, the t-test is,

$$t = \frac{|\bar{x}_{i.} - \bar{x}'_{i.}|}{s_{\bar{x}_{i.} - \bar{x}'_{i.}}}$$

where,

$$s_{\bar{x}_{i.} - \bar{x}'_{i.}} = s_1 \sqrt{\frac{1}{n_i} + \frac{1}{n_i}} = s_1 \sqrt{\frac{2}{4}}$$

$$= \sqrt{2.25}\sqrt{0.5} = 1.50 \times 0.7071 = 1.0607$$

In our example, the estimated t value for the means of the first pair of columns (cols. 1 and 2) is;

$$t = \frac{|23.25-28.75|}{1.0607} = 5.18$$

The tabulated t values (5 percent, 1 percent) for 6 d.f. are,

$$t_{6,\ 0.05} = 2.447 \qquad t_{6,\ 0.01} = 3.707$$

since,

$t > t_{6,\ 0.05}$ and $t > t_{6.\ 0.01}$ - highly significant etc.

2. Testing the significance of the means between the rows, the t-test is,

$$t = \frac{|\bar{x}_{.j} - \bar{x}'_{.j}|}{s_{\bar{x}_{.j} - \bar{x}'_{.j}}}$$

where,

$$s_{\bar{x}_{.j} - \bar{x}'_{.j}} = s\sqrt{\frac{1}{n_j} + \frac{1}{n_j}} = \sqrt{2.25}\sqrt{\frac{1}{3}+\frac{1}{3}}$$

$$= 1.50 \times 0.8165 = 1.22$$

The estimated value of t for the means of the first pair of rows (rows 1, 2) is,

$$t = \frac{|24.67-27.67|}{1.22} = 2.46$$

The estimated t is bigger than $t_{6,\ 0.05}$ but smaller than $t_{6,\ 0.01}$ - significant.

etc.

11.3 Three criteria of classification (one individual in each sub-class)

Let us assume that we have n sample values of a given normal variate, classified according to three criteria A, B, C, into a groups ($\ell = 1, 2, \ldots a$), b rows ($j = 1, 2, \ldots b$), and c columns ($i = 1, 2, \ldots c$). i.e. $n = abc$. Let $x_{\ell ji}$ be the value of the variate in the ℓ^{th} group, in the j^{th} row, in the i^{th} column.

Notation:

$\bar{x}_{\ell..}$: Mean in the ℓ^{th} group

$\bar{x}_{\ell j.}$: Mean in the j^{th} row within the ℓ^{th} group

$\bar{x}_{\ell . i}$: Mean in the i^{th} column within the ℓ^{th} group

$\bar{x}$: Overall mean, etc.

We can prove by a direct extension of the method already given for the simpler cases that, on the hypothesis of homogeneity and normality, the total sum of squares may be split into seven components ($7 = 2^3-1$).

Example

It is a common practice in the fish processing business to test the quality of the processed products by using smell-testing panels. The table below gives the sample data for three criteria of classification, i.e.

Factor A corresponds to three processed species i.e. Lates, Labeo, Hydrocyon ($\ell = 1, 2, 3$, $a = 3$).

Factor B corresponds to the brining times i.e. t_1, t_2, t_3 ($j = 1, 2, 3$, $b=3$).

Factor C corresponds to the smell-testing panel ($i = 1, 2, \ldots 9$, $c = 9$).

The values of variate x indicate the marks given on the quality of processed fish (maximum mark 60).

Sample values for three criteria of classification

Species (A)	Brining time (B)	Smell-testing panel (C) i= 1	2	3	4	5	6	7	8	9	Totals $T_{\ell j.}$	Remarks
Lates (ℓ=1)	j=1 t_1	15	15	27	30	34	21	20	30	51	243	
	j=2 t_2	17	16	19	24	56	41	25	20	50	268	
	j=3 t_3	16	17	20	26	56	46	28	19	40	268	
Total, $T_{1i.}$		48	48	66	80	146	108	73	69	141	$T_{1..}$ = 779	
Labeo (ℓ=2)	j=1 t_1	19	23	18	19	20	30	24	17	26	196	
	j=2 t_2	15	24	29	35	19	26	20	15	43	226	
	j=3 t_3	21	24	47	28	21	35	21	20	24	241	
Total, $T_{2i.}$		55	71	94	82	60	91	65	52	93	$T_{2..}$ = 663	
Hydrocyon (ℓ=3)	j=1 t_1	25	32	14	78	32	19	20	11	21	252	
	j=2 t_2	29	55	15	52	25	22	24	14	25	261	
	j=3 t_3	33	71	34	50	40	31	37	41	19	356	
Total, $T_{3i.}$		87	158	63	180	97	72	81	66	65	$T_{3..}$ = 869	
Columns totals $T_{..i}$		190	277	223	342	303	271	219	187	299	2311	→ Grand total
Rows totals $T_{.j.}$		j=1 : 243+196+252 =									691	2311
		j=2 : 268+226+261 =									755	
		j=3 : 268+241+356 =									865	

The table for the analysis of variance for three criteria of classification follows on page 77.

The meaning of interaction: In the following table we have spoken of interactions between the factors A, B and C. What do we mean by this term? In order to use a very simple example to explain the term let us move for a while to another field, agriculture. It is well known that fertilization affects the quality of plants. Let us now consider two plants, A and B. If A is twice as good as B and if this relationship holds true no matter what fertilization treatment we use then there is no interaction between the two types of treatment. But if A is twice as good as B when there is no fertilizer and B is better than A when fertilizer is present then the treatments do interact. Also, if A is much better than B in the presence of one fertilizer treatment but only a little better in the presence of some other fertilizer treatment this would be a form of interaction.

In our example we are dealing with more than two factors, we have interaction if factor A has one effect when an other factor B is present and a different effect when factor B is absent, then A and B are said to interact.

It is impossible to demonstrate that a significant interaction exists unless there is an independent estimate of the error of variance. It is necessary to test (F-test)

Analysis of variance for three criteria of classification

Source of variation	Sum of squares	d.f.	Estimate of variance	Sum of squares (algebraical transformations
Between groups (A)	$q_1=bc\sum_{\ell=1}^{a}(\bar{x}_{\ell\cdot\cdot}-\bar{\bar{x}})^2$	(a-1)	$s_1^2=q_1/a-1$	$q_1=\sum_{\ell=1}^{a}\frac{T_{\ell\cdot\cdot}^2}{n_{\ell\cdot\cdot}}-\frac{T^2}{n}$
Between rows (B)	$q_2=ca\sum_{j=1}^{b}(\bar{x}_{\cdot j\cdot}-\bar{\bar{x}})^2$	(b-1)	$s_2^2=q_2/b-1$	$q_2=\sum_{j=1}^{b}\frac{T_{\cdot j\cdot}^2}{n_{\cdot j\cdot}}-\frac{T^2}{n}$
Between columns (C)	$q_3=ab\sum_{i=1}^{c}(\bar{x}_{\cdot\cdot i}-\bar{\bar{x}})^2$	(c-1)	$s_3^2=q_3/c-1$	$q_3=\sum_{i=1}^{c}\frac{T_{\cdot\cdot i}^2}{n_{\cdot\cdot i}}-\frac{T^2}{n}$
Interaction (AB)	$q_4=c\sum_{\ell=1}^{a}\sum_{j=1}^{b}(\bar{x}_{\ell j\cdot}-\bar{x}_{\ell\cdot\cdot}-\bar{x}_{\cdot j\cdot}+\bar{\bar{x}})^2$	(a-1)(b-1)	$s_4^2=q_4/(a-1)(b-1)$	$q_4=\sum_{\ell=1}^{a}\sum_{j=1}^{b}\frac{T_{\ell j\cdot}^2}{n_{\ell j\cdot}}-q_1-q_2-\frac{T^2}{n}$
Interaction (BC)	$q_5=a\sum_{j=1}^{b}\sum_{i=1}^{c}(\bar{x}_{\cdot ij}-\bar{x}_{\cdot j\cdot}-\bar{x}_{\cdot\cdot i}+\bar{\bar{x}})^2$	(b-1)(c-1)	$s_5^2=q_5/(b-1)(c-1)$	$q_5=\sum_{j=1}^{b}\sum_{i=1}^{c}\frac{T_{\cdot ji}^2}{n_{\cdot ji}}-q_2-q_3-\frac{T^2}{n}$
Interaction (CA)	$q_6=b\sum_{i=1}^{c}\sum_{\ell=1}^{a}(\bar{x}_{\ell\cdot i}-\bar{x}_{\ell\cdot\cdot}-\bar{x}_{\cdot\cdot i}+\bar{\bar{x}})^2$	(c-1)(a-1)	$s_6^2=q_6/(c-1)(a-1)$	$q_6=\sum_{\ell=1}^{a}\sum_{i=1}^{c}\frac{T_{\ell\cdot i}^2}{n_{\ell\cdot i}}-q_1-q_3-\frac{T^2}{n}$
Residual (ABC)	$q_7=Q-(q_1+q_2+q_3+q_4+q_5+q_6)$	(a-1)(b-1)(c-1)	$s_7^2=q_7/(a-1)(b-1)(c-1)$	$q_7=Q-(q_1+q_2+q_3+q_4+q_5+q_6)$
Total	$Q=\sum_{\ell}\sum_{j}\sum_{i}(x_{\ell ji}-\bar{\bar{x}})^2$	(n-1)	-	$\sum_{\ell}\sum_{j}\sum_{i}x_{\ell ij}^2-\frac{T^2}{n}$

the estimates of σ^2 based on the "interaction" terms before testing the variation due to the individual factors separately. If none of these "interactions" when tested against the residual estimate of σ^2 is significant, we may pool the corresponding sums of squares, form a revised estimate of σ^2 based on more degrees of freedom and then proceed to test the estimates due to the individual factors; if, however, one or more of the estimates due to interactions is found to be significant, we have reason for suspecting the null hypothesis (remember, significant interactions do not necessarily imply real interactions between the factors concerned, e.g. apparent interaction may be due to experimental error).

Calculations: In order to simplify our calculations for the analysis of variances we shall regard first the factors A and B. Secondly the factors A and C, and thirdly the factors B and C. Finally, we shall complete the table of the analysis of variance.

a) The AxB table (species x brining time)

Species (A)		Brining time-B j=1	j=2	j=3	$T_{l\cdot\cdot}$
l=1	Lates	243	268	268	779
l=2	Labeo	196	226	241	663
l=3	Hydrocyon	252	261	356	869
$T_{\cdot j\cdot}$		691	755	865	2311

From the above table we shall calculate the variance of the individual factors, A, B and the interaction AxB.

1. Variance of A:

$$q_1 = \sum_{l=1}^{a} \frac{T_{l\cdot\cdot}^2}{n_{l\cdot\cdot}} - \frac{T^2}{n} = \{\frac{1}{27}(779^2+663^2+869^2) - \frac{(2311)^2}{81}\}$$

$$= \frac{1}{27}\times 1801571 - 65934.83 = 790.20$$

$$s_1^2 = q_1/a-1 = 790.2/2 = 395.10$$

2. Variance of B:

$$q_2 = \sum_{j=1}^{3} \frac{T_{\cdot j\cdot}^2}{n_{\cdot j\cdot}} - \frac{T^2}{n} = \{\frac{1}{27}(691^2+755^2+865^2) - 65934.83\} =$$

$$= \frac{1}{27}\times 1795731 - 65934.83 = 573.73$$

$$s_2^2 = q_2/b-1 = 573.73/2 = 286.86$$

3. Interaction AxB:

$$q_4 = \sum_{l=1}^{3}\sum_{j=1}^{3} \frac{T_{lj\cdot}^2}{n_{lj\cdot}} - q_1 - q_2 - \frac{T^2}{n}$$

$$= \{\frac{1}{9}(243^2+268^2+ \ldots +356^2) - (790.2+573.73+65934.83)\} =$$

$$= \frac{1}{9}\times 608631 - 67298.76 = 326.91$$

$$s_4^2 = q_4/(a-1)(b-1) = 326.91/4 = 81.73$$

b) The AxC table:

Species (A)		Smell-testing panel (C) i= 1	2	3	4	5	6	7	8	9	$T_{\ell\cdot\cdot}$
ℓ=1	Lates	48	48	66	80	146	108	73	69	141	779
ℓ=2	Labeo	55	71	94	82	60	91	65	52	93	663
ℓ=3	Hydrocyon	87	158	63	180	97	72	81	66	65	869
$T_{\cdot\cdot i}$		190	277	223	342	303	271	219	187	299	T=2311

1. Variance of C:

$$q_3 = \sum_{i=1}^{9} \frac{T^2_{\cdot\cdot i}}{n_{\cdot\cdot i}} - \frac{T^2}{n} = \{\tfrac{1}{9}(190^2+277^2+ \ldots +299^2)-65934.83\}$$

$$= \tfrac{1}{9}\times 617103-65934.83 = 2632.17$$

$$s_3^2 = q_3/8 = 2632.17/8 = 329.02$$

2. Interaction CxA:

$$q_6 = \sum_{\ell=1}^{3}\sum_{i=1}^{9} \frac{T_{\ell\cdot i}}{n_{\ell\cdot i}} - q_1 - q_3 - \frac{T^2}{n} = \{\tfrac{1}{3}(48^2+48^2+ \ldots +65^2)$$

$$-(790.2+2632.17+65934.83)\} = \tfrac{1}{3}\times 225797-68657.20$$

$$= 6608.47$$

$$s_6^2 = q_6/(a-1)(c-1) = 6608.47/16 = 413.03$$

c) The BxC table:

Brining time (B)	Smell-testing panel (C) i= 1	2	3	4	5	6	7	8	9	$T_{\cdot j\cdot}$
j=1	59	70	59	127	86	70	64	58	98	691
j=2	61	95	63	111	100	89	69	49	118	755
j=3	70	112	101	104	117	112	86	80	83	865

1. Interaction BxC:

$$q_5 = \sum_{j=1}^{3}\sum_{i=1}^{9} \frac{T_{\cdot ji}}{n_{\cdot ji}} - q_2 - q_3 - \frac{T^2}{n} = \{\tfrac{1}{3}(59^2+70^2+ \ldots +83^2)-$$

$$-(573.73+2632.17+65934.83)\} =$$

$$= \tfrac{1}{3}\times 210773-69140.73 = 1116.94$$

$$s_5^2 = q_5/(b-1)(c-1) = 1116.94/16 = 69.81$$

Finally,

$$Q = \sum_{\ell}\sum_{j}\sum_{i}(x_{\ell ij}-\bar{x})^2 = \sum_{\ell}\sum_{j}\sum_{i}x^2_{\ell ij} - \frac{T^2}{N} = (15^2+ \ldots +19^2)-65934.83$$

$$= 15306$$

and,

$$q_7 = Q-(q_1+q_2+q_3+q_4+q_5+q_6) =$$

$$= 15306-(790.2+573.73+2632.17+326.91+1116.94+6608.47)$$

$$= 3257.58$$

$$s_7^2 = q_7/(a-1)(b-1)(c-1) = 15306/32 = 101.80$$

Analysis of variance for three criteria of classification

Source of variation	Sum of squares	d.f.	Estimate of variance	F-values			Remarks
				Esti-mated	Tabulated 5%	Tabulated 1%	
Between species (A)	q_1=790.20	2	s_1^2=395.10	3.89*	3.30	5.43	Significant at 5% level
Between (B) brining time	q_2=573.73	2	s_2^2=286.86	2.83	3.30	5.34	Non-significant
Between panel (C)	q_3=2632.17	8	s_3^2=329.02	3.26**	2.25	3.12	Significant at 5% and 1% levs.
Interaction (AxB)	q_4=326.91	4	s_4^2=81.73	0.81	2.67	3.97	Non-significant
Interaction (BxC)	q_5=1116.94	16	s_5^2=69.81	0.69	1.97	2.62	Non-significant
Interaction (CxA)	q_6=6608.47	16	s_6^2=413.03	4.07**	1.97	2.62	Significant at 5% and 1% levs.
Residual (AxBxC)	q_7=3257.58	32	s_7^2=101.50	-	-	-	
Total	Q = 15306	80	-	-	-	-	

In the above table we have calculated the values of F in the respective cells in order to test the significance of interactions. As it is indicated in the table, the individual factor B is not significant. Also, since the interactions AxB and BxC are not significant, we may combine the corresponding sums of squares with that for residual to obtain a more accurate estimate of the assumed population variance, this is,

$$\frac{q_4+q_5+q_7}{4+16+32} = \frac{4701.43}{52} = 90.41$$

In the table below the new levels of significance are calculated by using the new estimate of σ^2.

Analysis of variance for three criteria of classification

Source of variation	Sum of squares	d.f.	Estimate of variance	F-values			Remarks
				Esti-mated	Tabulated 5%	Tabulated 1%	
Between species (A)	q_1=790.20	2	s_1^2=395.10	4.37*	3.18	5.06	Significant at 5% level
Between brining time (B)	q_2=573.73	2	s_2^2=286.86	3.17	3.18	5.06	Non significant
Between panel (C)	q_3=2632.17	8	s_3^2=329.02	3.64**	2.13	2.88	Significant at 1% and 5% levels
Interaction (CxA)	q_6=6608.47	16	s_6^2=413.03	4.57**	1.85	2.39	Significant at 1% and 5% levels
Residual	4701.43	52	s_7^2=90.41	-	-	-	
Total	Q = 15306	80	-	-	-	-	

Conclusions:

1. There are differences in the quality between the processed species.
2. Differences in the quality of fish products from brining time are not significant.
3. There are real differences between the panel members.
4. There is significant interaction between A and C i.e. the kind of fish has a different effect on the obtained marks concerning the quality of the processed products.

11.4 Analysis of covariance

The analysis of covariance is a technique that combines the analysis of variance and regression[1/]. Suppose that for each survey unit we have a pair of variates x, y and that we are interested in the effect of classification on the relationship between x and y, as expressed by their covariance. This leads to the analysis of covariance.

11.4.1 One-way classification

In one way classification the survey units are grouped in k classes (one factor), and for each unit we have a pair of variates x_{ij}, y_{ij}.

Notation:

i : stands for a given class ($i = 1, 2, \ldots k$)

j : stands for a given unit.

n_i : sample size in i^{th} class ($j = 1, 2, \ldots n_i$)

$n = \sum_i n_i$: Overall sample size.

$T_{xi} = \sum_j x_{ij}$: Sample total of x in the i^{th} class

$T_{yi} = \sum_j y_{ij}$: Sample total of y in the i^{th} class.

Within classes:

$$C_{xyw} = \sum_j (x_{ij}-\bar{x}_i)(y_{ij}-\bar{y}_i) = \sum_j x_{ij}y_{ij} - \frac{(\sum_j x_{ij})(\sum_j y_{ij})}{n_i}$$

$$C_{xxw} = \sum_j (x_{ij}-\bar{x}_i)^2 = \sum_j x_{ij}^2 - \frac{(\sum_j x_{ij})^2}{n_i}$$

$$C_{yyw} = \sum_j (y_{ij}-\bar{y}_i)^2 = \sum_j y_{ij}^2 - \frac{(\sum_j y_{ij})^2}{n_i}$$

Between classes:

$$C_{xyb} = \sum_i \frac{(T_{xi})(T_{yi})}{n_i} - \frac{(T_x)(T_y)}{n} \quad , (T_x = \sum_i T_{xi}, \quad T_y = \sum_i T_{yi})$$

1/ See Chapter 13.

$$C_{xxb} = \sum_i \frac{T_{xi}^2}{n_i} - \frac{T_x^2}{n}$$

$$C_{yyb} = \sum_i \frac{T_{yi}^2}{n_i} - \frac{T_y^2}{n}$$

Total:

$$C_{xyT} = \sum_{ij}\sum (x_{ij}-\bar{x})(y_{ij}-\bar{y}) = \sum_{ij}\sum x_{ij}y_{ij} - \frac{(T_x)(T_y)}{n}$$

$$C_{xxT} = \sum_{ij}\sum (x_{ij}-\bar{x})^2 = \sum_{ij}\sum x_{ij}^2 - \frac{T_x^2}{n}$$

$$C_{yyT} = \sum_{ij}\sum (y_{ij}-\bar{y})^2 = \sum_{ij}\sum y_{ij}^2 - \frac{T_y^2}{n}$$

Example

The data in the table below are the results of an experiment at lake L, aiming to study the effect of the "barrier" factor on the catchability level of gillnet units of various kinds. Specifically, in the experiment three kinds of gillnet units (one factor of classification) were set in randomly selected places (fixed points) and fished without being removed for a number of days. The variate x_{ij} expresses the number of fish caught per gillnet unit on the first day of the experiment and the variate y_{ij} expresses the number of fish caught per gillnet unit on the third day of the experiment.

- number of fish -

Kind of gillnets used						
A (i=1)		B (i=2)		C (i=3)		
x	y	x	y	x	y	($\sum_{ij}\sum x_{ij}y_{ij}$=3173)
12	5	7	0	17	12	
9	0	7	1	14	9	($\sum_{ij}\sum x_{ij}^2$=4796)
6	1	8	2	12	17	
15	7	9	0	10	4	($\sum_{ij}\sum y_{ij}^2$=2714)
20	10	19	17	22	22	
7	3	9	3	17	11	($\frac{T_x^2}{n}$=4130)
11	12	20	13	13	4	
7	0	9	8	13	15	($\frac{T_y^2}{n}$=1470)
12	7	6	0	8	0	
4	0	16	8	13	19	$\bar{x}$ = 11.73
T_{xi}=103 T_{yi}=45		T_{xi}=110 T_{yi}=52		T_{xi}=139 T_{yi}=113		T_x=352, $\bar{x}$=11.73
$\bar{x}_i$=10.3 $\bar{y}_i$=4.5		$\bar{x}_i$=11.0 $\bar{y}_i$=5.2		$\bar{x}_i$=13.9 $\bar{y}_i$=11.3		T_y=210, $\bar{y}$=7.00

Analysis of sum of squares and products

Source of variation	d.f.	Sum of squares for: x	Sum of squares for: y	Products xy	
Total	n-1 =29	C_{xxT}=666	C_{yyT}=1244	C_{xyT}=709	
Between columns	k-1 =2	C_{xxb}=73	C_{yyb}=279	C_{xyb}=142.20	
Error (within columns)	n-k =27	$E_{xx}=C_{xxT}-C_{xxb}$=593	$E_{yy}=C_{yyT}-C_{yyb}$=965	$E_{xy}=C_{xyT}-C_{xyb}$=566.80	$s^2_y=\frac{E_{yy}}{n-k}$=35.74
Reduction due to regression	1		E^2_{xy}/E_{xx}= 541.76		
Deviations from regression	n-k-1 =26		E_1=965-541.76=423.24	Mean square→	$s^2_{y\cdot x}=\frac{E_1}{n-k-1}$ = 16.28

In the analysis of covariance (one criterion of classification) the interest is concentrated in the test of significance relating to the adjusted means. In such a case the following four steps should be taken:

1. Compute the value of the regression coefficient

$$b = E_{xy}/E_{xx} = \frac{566.80}{593} = 0.9558$$

2. Calculate the values of the adjusted means. The adjusted means are given by:

$$\bar{y}'_i = \bar{y}_i - b(\bar{x}_i - \bar{x})$$

and,

$$\bar{y}'_1 = 4.5-0.9558(10.3-11.73) = 5.87$$

$$\bar{y}'_2 = 5.2-0.9558(11.0-11.73) = 5.90$$

$$\bar{y}'_3 = 11.3-0.9558(13.9-11.73) = 9.23$$

3. For the test of significance or confidence limits relating to the adjusted means the error variance is derived from the mean square $s^2_{y\cdot x}$. Algebraically the difference between two adjusted means is given by:

$$D = (\bar{y}'_i - \bar{y}'_{i'}) = (\bar{y}_i - \bar{y}_{i'}) - b(\bar{x}_i - \bar{x}_{i'})$$

Since D is different for each pair of classes, the average variance of D ($\bar{s}^2_D$) is used to compute the error variance for any comparison among the classes,

$$\bar{s}^2_D = \frac{2s'^2}{n_i}$$

where,

$$s'^2 = s^2_{y\cdot x}\left(1+\frac{t_{xx}}{E_{xx}}\right) \quad , \quad t_{xx} = C_{xxb}/k-1$$

and,

$$s'^2 = 16.28\left(1+\frac{36.5}{593}\right) = 17.28$$

$$\bar{s}_D^2 = \frac{2\times 17.28}{10} = 3.456$$

$$\bar{s}_D = 1.86$$

4. The t-values for the pair of adjusted means are,

$(t_{a=0.05,\ \nu=26} = 2.06)$

$$(A, B),\ t = \frac{|\bar{y}'_1-\bar{y}'_2|}{\bar{s}_D} = \frac{|5.87-5.90|}{1.86}$$

$$= 0.02, \text{ not significant}$$

$$(B, C),\ t = \frac{|\bar{y}'_1-\bar{y}'_3|}{\bar{s}_D} = \frac{|5.87-9.23|}{1.86}$$

$$= 1.81, \text{ not significant}$$

$$(A, C),\ t = \frac{|\bar{y}'_2-\bar{y}'_3|}{\bar{s}_D} = \frac{|5.90-9.23|}{1.80}$$

$$= 1.85, \text{ not significant}$$

Conclusion: The kind of gillnets used does not affect the existing relationship between x_{ij} and y_{ij} or the various kinds of nets used respond in the same way to the "barrier" effect.

The relative efficiency (RE) of the adjusted means relative to the unadjusted means is estimated by the ratio of the corresponding effective mean squares:

$$RE = \frac{s_y^2}{s'^2} = \frac{35.74}{17.28} = 2.07$$

Covariance with 10 replicates per class gives nearly as precise estimates as the unadjusted means with 21 replicates.

The F-test of the adjusted means: The F-test can also be used to test the null hypothesis Ho that all the population means are equal - that there are no differences among the adjusted means. Because the Ho model contains no treatment effects, in the analysis of sums of squares and products for the Ho model, the error line is the sum of the error and treatments line in the original model. The estimated F-value is given by:

$$F = \frac{s_A^2}{s_{y\cdot x}^2}, \quad (F_{\nu_1=k-1,\ \nu_2=n-k-1,\ a=\ldots} \text{ - tabulated})$$

where,

$$s_A^2 = \frac{1}{k-1}\left(C_{yyT} - \frac{C_{xyT}^2}{C_{xxT}} - E_1\right)$$

$$= \frac{1}{2}\left(1244 - \frac{(709)^2}{666} - 423.24\right) = 32.99$$

and,

$$F = \frac{32.99}{16.28} = 2.03, \quad F_{\nu_1=2,\ \nu_2=26} = 3.37\ (5\%),\ 5.53\ (1\%)$$

Conclusion: The hypothesis that there are no differences between the adjusted means is valid.

11.4.2 Two-way classification

In the analysis of covariance (two-way classification) we have a two-factor experiment and for each survey unit we have a pair of variates x, y. In this technique the practical interest is concentrated on estimating the regression coefficients, testing the linearity of the estimated regression equations and, if they differ in slope, elevation or in residual variance[1/].

12. EXPERIMENTAL DESIGNS

12.1 Introduction

The analysis of variance and covariance are statistical methods aimed at extracting as much information as possible from the results of an experiment. It is important, of course, that the experiment should be properly designed in the first place and a great deal of work has been done in developing efficient experimental designs and procedures[2/]. In this chapter, after a brief discussion of a few of the general principles involved in the design of experiments, we present some of the designs which are of interest in the field of applied fishery researches in inland waters.

12.2 Randomization, replication and selection of treatments

The design of the experiment is an extremely important step because errors made in the design can invalidate the results of the entire experiment. Several principles apply to all experimental designs:

1. Randomization: In the applied fishery researches one can think of many factors which might affect the outcome of experimental units; light intensity, variation in temperature, etc. Such factors may change with time, with location, with a change from one fish species to another and so on. As indicated, we are often dealing with unknown factors which vary in unknown ways. How then, can we minimize these sources of bias? The answer lies in randomization, or the process of determining which experimental units are to receive a given treatment by a purely random fashion.
2. Replication: Randomization by itself is not necessarily sufficient to yield a valid experiment. Replication of the experiment or experimental units is also necessary for the application of any method that obtains its measure of variability directly from the data because at least two observations are needed to measure variation.

1/ For the application of the technique see, G.P. Bazigos, Efficiency of Different Sampling Methods for Large Scale Biological and Fishery Statistical Sample Surveys at Large African Lakes: 1. Cove-Rotenone Sample Survey at Kariba Lake, St.E./1, UNDP/SF/ZAM/11, April 1972.

2/ R.A. Fisher, The Design of Experiments (5th edition 1949), Cochran and Cox, Experimental Designs, 1950.

3. Selection of treatments: Treatments are the various states of any factor to be studied. The information to be gleaned from an experiment may often be greatly increased by careful consideration at the outset of the treatments to be applied. A thorough knowledge of the experimental material and some idea about the likely effects of the treatments are very helpful in this regard. In applying the treatments, one should remember that the first and one of the most important precautions to take in carrying out an experiment is to label properly and adequately all the experimental units.

12.3 The latin square

The latin square is a device for controlling two sources of error at once.

A	B	C	D
B	A	D	C
C	D	B	A
D	C	A	B

As indicated in the above latin square the treatments (A, B, C, D) are so allocated among the cells that no treatment occurs more than once in any one row or any one column. In a latin square of α rows and columns, let R_j denote the sum of the j^{th} row, C_i the sum of the i^{th} column and T_k the sum of the k^{th} treatment for the variable x ($i, j, k = 1, 2, \ldots \alpha$). Let G be the grand total. The analysis of variance is,

Table of analysis of variance

Variability	Sum of squares	d.f.	Estimate of variance
1.Rows	$\zeta_1 = \sum_j \frac{R_j^2}{a} - \frac{G^2}{a^2}$	a-1	$s_1^2 = \zeta_1 / a-1$
2.Columns	$\zeta_2 = \sum_i \frac{C_i^2}{a} - \frac{G^2}{a^2}$	a-1	$s_2^2 = \zeta_2 / a-1$
3.Treatments	$\zeta_3 = \sum_k \frac{T_k^2}{a} - \frac{G^2}{a^2}$	a-1	$s_3^2 = \zeta_3 / a-1$
4.Error	$\zeta_4 = \zeta - (\zeta_1 + \zeta_2 + \zeta_3)$	(a-1)(a-2)	$s_4^2 = \zeta_4 / (a-1)(a-2)$
5.Total	$\zeta = \sum_j \sum_k x_{jk}^2 - \frac{G^2}{a^2}$	a^2-1	-

It should be noted that for a large number of treatments it is often difficult to arrange for the right number of rows and columns, and for a small number there are not many degrees of freedom left for error.

Example

Assuming four kinds of longlines must be tested in one day and that a fisherman must fish at least three hours each longline to get an adequate measure of the

catchability level of the fishing gear, the 12-hour working day will be divided into four three-hour periods. However, now a third factor has entered the experiment, the variation in light intensity which might affect the obtained catches. We have then three factors: fishing gear, fishermen and time periods. But it is impossible to obtain observations for all combinations of all levels since, for example, all the fishermen cannot fish the same longline during the first time period. The difficulty is met by setting up the experiment so that all three factors appear at the same number of levels. Thus, since there are four longlines and four fishing time periods, we should use four fishermen in the experiment. In the latin square the letters A, B, C, D correspond to the four longlines, the rows correspond to fishermen and the columns to fishing time, then the second fisherman fishes the first longline (A) during the third time period (see table below) etc.

Table of latin square (x : number of fish caught)

Fishermen	Fishing time i=1	i=2	i=3	i=4	R_j
j=1	A:14	B:10	C:9	D:8	41
j=2	C:15	D:17	A:12	B:10	54
j=3	B:11	C:11	D:9	A:12	43
j=4	D:13	A:10	B:10	C:10	43
C_i	53	48	40	40	T=181

Table of longlines

A	B	C	D
14	10	9	8
12	10	15	17
12	11	11	9
10	10	10	13
T_k: 48	41	45	47

Table of analysis of variance

Variability	Sum of squares	d.f.	Estimate of variance	F-values Calculated	F-values Tabulated 5%	F-values Tabulated 1%	Remarks
1.Rows (fishermen)	ζ_1=2073.75-2047.56 =26.19	3	8.73	2.24	4.76	9.78	Non-significant
2.Columns (time period)	ζ_2=2078.25-2047.56 =30.69	3	10.23	2.62	4.76	9.78	Non-significant
3.Treatments (longlines)	ζ_3=2054.75-2047.56 =7.19	3	2.40	0.62	4.76	9.78	Non-significant
4.Error	ζ_4=87.44-67.07 =23.37	6	3.90	-	-	-	
Total	ζ=2135-2047.56 =87.44	15	-	-	-	-	

Conclusions: In our hypothetical example, the latin square indicates no difference in the catch level from the effect of the factors under consideration.

12.4 Factorial experiments

When two or more factors (each may be at two or more levels) are tested in all possible combinations, the resulting treatments are said to be factorial. Such treatment combinations make it possible to learn of a differential effect of one factor at two or more levels of a second factor, interaction. Even if interactions do not occur in such an experiment, the results are more widely applicable because the main treatments have been tried over a wider range of conditions.

Example

In cage-culture of fish at lake L it was desired to test the effect of various levels of a single ingredient in artificially prepared fish feed on the growth of three fish species. The three species were designated A, B, C and the levels of the ingredient by 0, 100 and 200 gr per kg of feed. The nine possible treatment combinations (three species times three levels of ingredient, 3^2) were assigned at random to nine fish-cases in each of three clusters (blocks). The effect of treatment was evaluated on the basis of average monthly growth over the next three months (x : expresses gr per month).

Block data

Clusters I

C_2 7	B_2 8	C_0 27
A_0 35	B_0 14	C_1 10
B_1 11	A_2 14	A_1 7

II

B_2 13	C_2 10	A_1 8
B_1 8	A_2 4	C_0 33
A_0 30	C_1 15	B_0 25

III

C_1 9	C_2 11	A_0 27
B_1 9	A_2 7	A_1 18
B_0 19	C_0 29	B_2 5

A, B and C = Species

Subscripts = ingredient level (0 = 0, 1 = 100, 2 = 200).

Table of cluster data

Species	Ingredient level	Blocks I(i=1)	Blocks II(i=2)	Blocks III(i=3)	Ingredient sub-totals $T_{\ell j\cdot}$	Species totals
A (ℓ=1)	j=1 0	35	30	27	92	
	j=2 1	7	8	18	33	
	j=3 2	14	4	7	25	
	Cluster sub-totals $T_{\ell\cdot i}$	56	42	52		$T_{\ell\cdot\cdot}$=150
B (ℓ=2)	j=1 0	14	25	19	58	
	j=2 1	11	8	9	28	
	j=3 2	8	13	5	26	
	Cluster sub-totals $T_{\ell\cdot i}$	33	46	33		$T_{\ell\cdot\cdot}$=112
C (ℓ=3)	j=1 0	27	33	29	89	
	j=2 1	10	15	9	34	
	j=3 2	7	10	11	28	
	Cluster sub-totals $T_{\ell\cdot i}$	44	58	49		$T_{\ell\cdot\cdot}$=151
All species	j=1 0	76	88	75	239	
	j=2 1	28	31	36	95	
	j=3 2	29	27	23	79	
	Cluster sub-totals $T_{\cdot\cdot i}$	133	146	134		T = 413

Calculations

a) The preliminary analysis of the nine treatment combinations is made as though this were a straight randomized block-design (temporarily ignoring their factorial nature).

1. Overall sum of squares:

$$\lambda = \sum_{\ell}\sum_{j}\sum_{i} x_{\ell ji}^2 - \frac{T^2}{n} = (35^2+7^2+\ldots+11^2) - \frac{413^2}{27}$$

$$= 8593-6317.37 = 2275.63$$

2. Variance of blocks:

$$\lambda_1 = \sum_{\ell=1}^{3} \frac{T_{\cdot\cdot i}^2}{n_{\cdot\cdot i}} - \frac{T^2}{n} = \{\tfrac{1}{9}(133^2+146^2+134^2) - 6317.37\}$$

$$= 6329-6317.37 = 11.63$$

$$s_1^2 = 11.63/2 = 5.82$$

3. Variance of treatment combinations:

$$\lambda_2 = \sum_{j=1}^{3} \frac{T_{\ell j \cdot}^2}{n_{\ell j \cdot}} - \frac{T^2}{n} = \{\frac{1}{3}(92^2+33^2+ \ldots +28^2)-6317.37\}$$

$$= 8287.67-6317.37 = 1970.30$$

$$s_2^2 = \lambda_2/8 = 1970.30/8 = 246.29$$

4. Error:

$$\lambda_6 = \lambda-\lambda_1-\lambda_2 = 2275.63-(11.63+1970.30) = 293.64$$

$$s_6^2 = \lambda_6/16 = 293.64/16 = 18.35$$

Tabulating the calculated variances we get:

Source of variation	d.f.	Sum of squares	Estimate of variance	F-values			Remarks
				Calculated	Tabulated 5%	Tabulated 1%	
Blocks	2	λ_1=11.63	s_1^2=5.82	0.32	3.63	6.23	Non-significant
Treatments	8	λ_2=1970.30	s_2^2=246.29	13.41**	2.59	3.89	Highly significant
Error	16	λ_6=293.64	s_6^2= 18.35	-	-	-	
Totals	26	λ=2275.57	-	-	-	-	

Conclusions: From the above table it is indicated that the difference between the treatment combinations is highly significant. The next step is to analyze the components of the nine treatment combinations. How do the species compare? What is the effect of the ingredient levels? And does the ingredient affect all species the same way?

b) Table species x ingredient levels

Species	Ingredient levels			Species totals $T_{\ell \cdot \cdot}$
	0	1	2	
A	92	33	25	150
B ($T_{\ell j \cdot}$)	58	28	26	112
C	89	34	28	151
Ingredient level totals $T_{\cdot j \cdot}$	239	95	79	(T)413

1. Variance of species

$$\lambda_3 = \sum_{\ell=1}^{3} \frac{T_{\ell \cdot \cdot}^2}{n_{\ell \cdot \cdot}} - \frac{T^2}{n} = \{\frac{1}{9}(150^2+112^2+151^2)-6317.37\} = 109.85$$

$$s_3^2 = \lambda_3/2 = 109.85/2 = 54.93$$

2. **Variance of ingredient levels**

$$\lambda_4 = \sum_j \frac{T^2_{\cdot j \cdot}}{n_{\cdot j \cdot}} - \frac{T^2}{n} = \{\tfrac{1}{9}(239^2+95^2+79^2)-6317.37\} =$$

$$= 8043.00-6317.37 = 1725.63$$

$$s_4^2 = \lambda_4/2 = 1725.63/2 = 862.82$$

3. Species x ingredient level interaction

$$\lambda_5 = \sum_{\ell j} \frac{T^2_{\ell j \cdot}}{n_{\ell j \cdot}} - \lambda_3 - \lambda_4 - \frac{T^2}{n} = \{\tfrac{1}{3}(92^2+33^2+ \ldots +28^2)-$$

$$-(109.85+1725.63+6317.37)\} = 134.82$$

$$s_5^2 = \lambda_5/4 = 134.82/4 = 33.71$$

Tabulating the calculated variances of the above (a) and (b) we get:

Source of variation	d.f.	Sum of squares	Estimate of variance	F-values			Remarks
				Calculated	Tabulated 5%	Tabulated 1%	
Blocks	2	λ_1=11.63	s_1^2=5.82	0.32	3.63	6.23	Non-significant
Treatments	8	λ_2=1970.30	s_2^2=246.29	13.41**	2.59	3.89	Highly significant
Species	2	λ_3=109.85	s_3^2=54.93	2.99	3.63	6.23	Non-significant
Ingredient levels	2	λ_4=1725.63	s_4^2=862.82	46.99**	3.63	6.23	Highly significant
Species × ingredient levels	4	λ_5=134.82	s_5^2=33.71	1.88	3.01	4.77	Non-significant
Error	16	λ_6=293.64	s_6^2=18.35	-	-	-	-
Total	26	λ=2275.57	-	-	-	-	-

Conclusions: The additional element portrayed by the above tabulation is that there are significant differences in the ingredient levels.

13. CORRELATION AND REGRESSION

13.1 Introduction

Relationships are the fundamentals from which all science is built. Specifically, in some problems the several variables are studied simultaneously to see how they are interrelated, in others there is one particular variable of interest and the remaining variables are studied for their possible aid in throwing light on this particular variable. These two classes of problems are usually associated with the names of correlation and regression respectively.

13.2 Correlation (bivariate distributions)

In the field of applied fishery studies it frequently happens that the size of two associated variables is related (bivariate distribution). As an example, the length of

a fish and its weight tend to be related. When a fish is long it is also heavy. When a fish is short the weight of the fish is low. Variables exhibiting such relationships are said to be correlated.

13.2.1 Linear correlation

When stating the degree of relationship between two variables, it is convenient to be able to employ concise numerical terms which are independent of the units of the original data and to express the degree of relationship between the two variables. When the relationship between two variables may be treated as approximately linear[1] the calculated (estimated) measure of relationship is called the coefficient of linear correlation and is denoted by the letter r. The linear coefficient of correlation is a number varying from +1, through zero, to -1. Specifically the sign indicates whether the slope of the line of relationship[1] is positive or negative, while the magnitude of the coefficient indicates the degree of association. When there is absolutely no relationship between the variables r is zero[2].

The investigation of the relationship between two associated variables x, y, based upon a set of n pairs of measurements $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, usually begins by graphing the data as n points in the x, y plane. Since we confine ourselves to two dimensions, we make a scatter-diagram. In the x, y plane, the abscissa being values of one of the variates, the ordinate those of the other variate. For every pair of values (x_i, y_i) there will be a corresponding point in the plane of the axes. If we plot these points we have a scatter-diagram (see Figure 13.2.1.1).

The linear correlation coefficient is defined by the following formula,

$$r = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\{\sum_i (x_i-\bar{x})^2\}\{\sum_i (y_i-\bar{y})^2\}}} = \frac{\sum \dot{x}\dot{y}}{\sqrt{\sum \dot{x}^2 \sum \dot{y}^2}}$$

where,

$$\dot{x} = x_i-\bar{x} \quad , \quad \dot{y} = y_i-\bar{y}$$

or,

$$r = \frac{\sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n}}{\sqrt{\left(\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}\right)\left(\sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n}\right)}}$$

or,

$$r = \frac{n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{\sqrt{\{n\sum_i x_i^2 - (\sum_i x_i)^2\}\{n\sum_i y_i^2 - (\sum_i y_i)^2\}}}$$

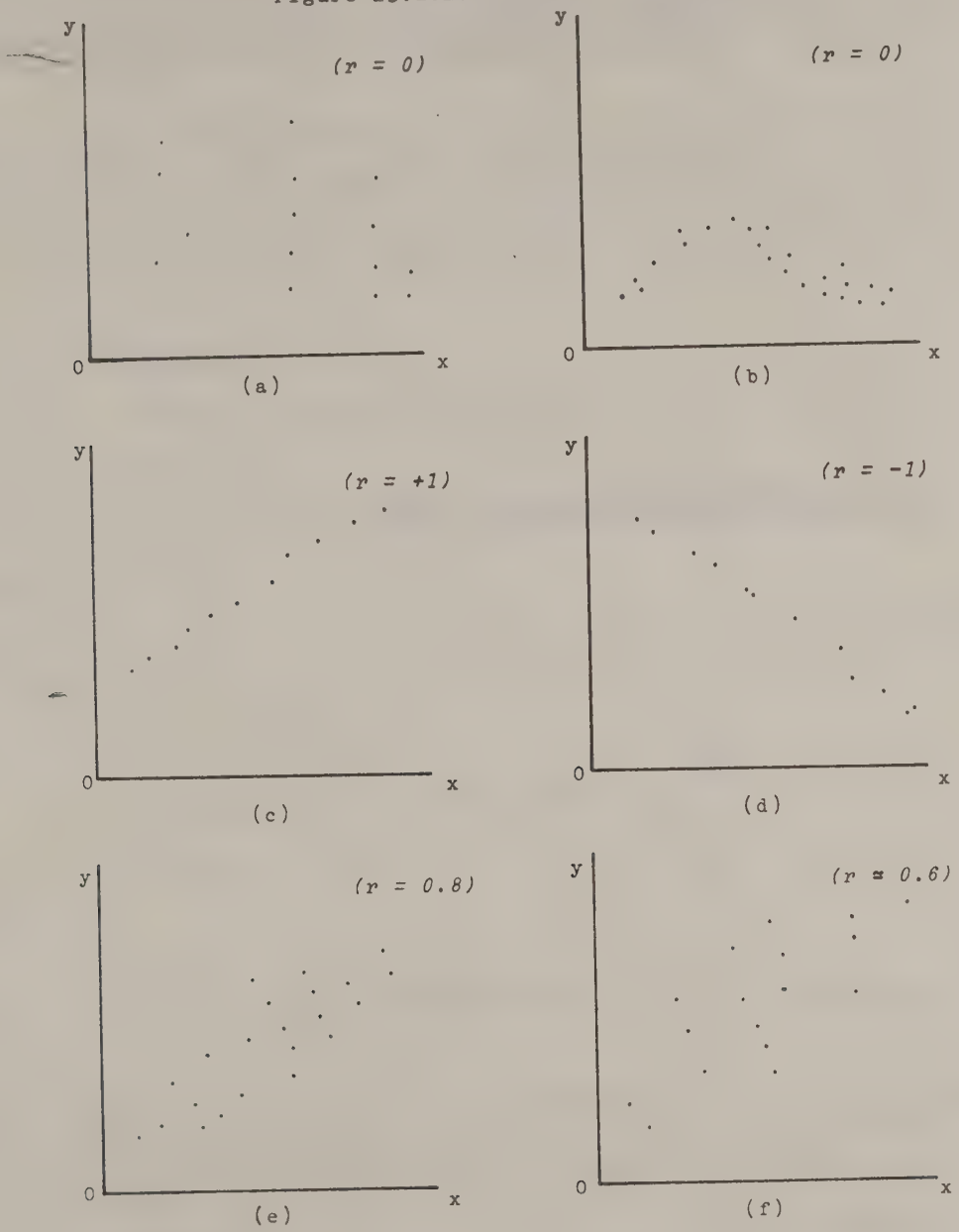
Example

In a pilot survey at a man made lake L two approaches were used in order to provide estimates of the total number of fishing boats by fishing site i.e. aerial approach (variate x) and water approach (variate y). Prepare the respective scatter diagram and calculate the value of r.

1/ See also section 13.3.

2/ Any value of r outside of the range, $-1 \leq r \leq +1$, means an error in calculation.

Figure 13.2.1.1 Scatter diagrams

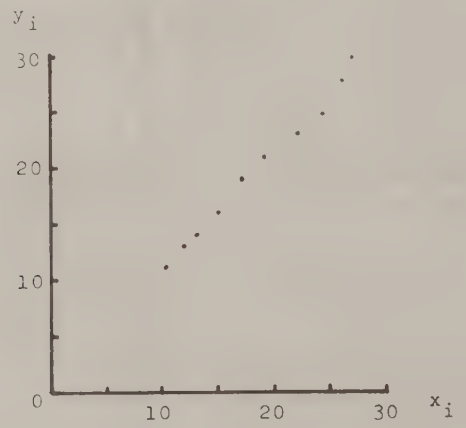


Note: A value of r = ±1 indicates perfect correlation. A value of r = 0 indicates no correlation. A plus value of r indicates that as one variable increases the other does too. A negative value of r indicates that as one variable increases the other decreases.

Remember that the correlation coefficient tells us how closely two variables are related. It does not imply a cause and effect relationship.

Fishing site	x_i (boats seen)	y_i (boats counted)	Remarks
1	10	11	(n = 10)
2	12	13	($\sum x_i y_i = 4054$)
3	13	14	
4	15	16	($\sum x_i^2 = 3753$)
5	17	19	
6	19	21	($\sum y_i^2 = 4382$)
7	22	23	$\{(\sum x_i)^2 = 34225\}$
8	24	25	
9	26	28	$\{(\sum y_i)^2 = 40000\}$
10	27	30	
	185	200	

$$r = \frac{10\times4054-185\times200}{\sqrt{(10\times3753-34225)(10\times4382-40000)}} = \frac{3540}{3553} = 0.9963$$



Is r significant? To test the significance of the calculated r we assume that there is no correlation between the two variables. Then using the t-test we calculate the probability of obtaining a value of r as large as the calculated value if there is no correlation (that is, if r is actually 0). The t value is given by the formula,

$$t = r\sqrt{\frac{(n-2)}{1-r^2}}$$

In the example given above,

$$t = 0.9963\sqrt{\frac{8}{1-0.9963^2}} = 0.9963\sqrt{\frac{8}{0.0074}} = 32.7581$$

The tabulated t values for 0.05 and 0.01 are, respectively,

$t_{8,\ 0.05} = 2.306$ $\quad$ $t_{8,\ 0.01} = 3.355$

As the calculated value of t is much larger than the tabulated ones we can say that there is less than one percent chance of obtaining so large a value of r if there is actually no correlation between the two variables. In other words, by using a shortcut test, we reject the hypothesis that the two variables are not correlated because the calculated t value is much higher than the tabulated t values at 0.05 and 0.01.

Further notes on the reliability of r: If ρ is the surveyed parameter (correlation coefficient) in a bivariate normal distribution then r may be used to estimate the value of ρ, just as the sample mean $\bar{x}$ is used to estimate the population mean m. However, the frequency function of r is decidedly non-normal for large values of ρ, and depends only upon the parameters ρ and n, where n is the number of points in the scatter diagram. It has been proved that there exists a simple change of variable which transforms the complicated distribution of r into an approximately normal distribution,

$$z = \frac{1}{2}\log_e\frac{1+r}{1-r}$$

and

$$\sigma_z = \frac{1}{\sqrt{n-3}}$$

The transformer[1/] z can be used both to test a hypothesis about ρ and to establish confidence intervals for ρ.

Example

The calculated value of r based on a sample of 38 pairs of numbers was r = 0.60. Test the hypothesis that $\rho = 0$. If $\rho \neq 0$, estimate the confidence limits for ρ (P = 95 percent).

1. $z = \frac{1}{2}\log_e\frac{1+0.60}{1-0.60} = 0.6931$

$$\sigma_z = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{38-3}} = \frac{1}{5.9161} = 0.1690$$

$$t = \frac{0.6931}{0.1690} = 4.1012$$

$t > t_{a=0.05}$, $4.1012 > 1.96$

Conclusion: The hypothesis that Ho : $\rho = 0$ is discredited.

2. $z - 1.96\sigma_z < Z < z + 1.96\sigma_z$

$0.6931 - 1.96 \times 0.1690 < Z < 0.6931 + 1.96 \times 0.1690$

$0.3619 < Z < 1.0243$

and,

$0.3471 < \rho < 0.7718$

Example

The estimated linear correlation coefficients, based on two independent samples of sizes $n_1 = 67$ and $n_2 = 84$, were respectively $r_1 = 0.70$ and $r_2 = 0.76$. Test the hypothesis that the two samples have been selected from the same population.

1/ For tables giving a) the values of z for r from .00 to 1.00, and b) the values of r for z from 0 to 5, see Appendix II.

$$n_1 = 67 \ , \quad r_1 = 0.70$$

$$z_1 = \frac{1}{2}\log_e\frac{1+0.70}{1-0.70} = 0.8673$$

$$n_2 = 84 \ , \quad r_2 = 0.76$$

$$z_2 = \frac{1}{2}\log_e\frac{1+0.76}{1-0.76} = 0.9962$$

$$\sigma_{z_1-z_2} = \sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}} = 0.1670$$

$$t = \frac{|0.9962-0.8673|}{0.1670} = 0.7719$$

Conclusion: Since $t<t_{a=0.05}$, $0.7719<1.96$, the hypothesis is valid.

13.2.2 Rank correlation (non-parametric methods)

Spearman's rank correlation coefficient r_s is used to provide a quick estimate of the level of concordance between two variables, the formula for which is,

$$r_s = 1-\frac{6\sum_i d_i^2}{n(n-1)}$$

where d_i refers to the differences in rank between paired values of the individuals[1/].

Example

The table below gives the number of houses seen (variate x) and boats seen (variate y) per fishing site in an aerial flight which covered a limnological zone (lake L). Calculate the value of r_s.

Fishing site	Houses seen x	Boats seen y	Rank order x	Rank order y	d_i	d_i^2
1	10	28	1	1	0	0
2	12	35	2	3	-1	1
3	13	30	3	2	1	1
4	14	36	4	4	0	0
5	16	40	5	6	-1	1
6	17	37	6	5	1	1
7	19	42	7	7	0	0
8	20	48	8	8	0	0
9	22	52	9	10	-1	1
10	25	50	10	9	1	1
						6

$$r_s = 1-\frac{6\times6}{90} = 1-\frac{36}{90} = 0.60$$

The values of r_s can range in samples from +1 (complete concordance) to -1 (complete disagreement).

13.2.3 Correlation of grouped data

When the number of pairs of items to be correlated is large, time is saved if the data are grouped before calculations are undertaken. The following table (Table 13.2.3.1) is a correlation table, showing the relationship between number of hooks owned and gillnets owned by the FEU's in a big fishing site at lake L. The table resembles

1/ Another measure of degree of concordance, closely related to r_s is Kendall's τ.

a scatter diagram except that each point, instead of being plotted exactly, is merely entered in the appropriate cell.

Table 13.2.3.1 Joint frequency distribution of the number of hooks owned (x) and number of gillnet units owned (y) by the FEU's of a large fishing site at lake L

No. of gillnet units owned (y_j)			No. of hooks owned (x_i)					f_y	$y'f_y$	$\sum x'y'f_{xy}$
			10 (i=1)	20 (i=2)	30 (i=3)	40 (i=4)	50 (i=5)			
		y'	x':-2	-1	0	1	2			
j=1	2	-3	8	12	5			25	-75	84
j=2	3	-2	10	18	8	10	5	51	-102	36
j=3	4	-1	7	15	17	8	8	55	-55	5
j=4	5	0	5	3	23	12	12	55	0	0
j=5	6	1		2	17	20	10	49	49	38
j=6	7	2				10	5	15	30	40
		f_x	30	50	70	60	40	n=250	-153	203
		$x'f_x$	-60	-50	0	60	80	⟶	+30	

Table 13.2.3.2 Calculation of the linear correlation coefficient in the above table

$\sum x'f_x = 30$	$\sum y'f_y = -153$	
(1) $\sum x'^2 f_x = 390$	(1) $\sum y'^2 f_y = 593$	(1) $\sum x'y'f_{xy} = 203$
(2) $(\sum x'f_x)^2/n = 3.60$	(2) $(\sum y'f_y)^2/n = 93.64$	(2) $(\sum x'f_x)(\sum y'f_y)/n = -18.36$
(1)-(2) = $\sum \dot{x}^2 = 386.40$	(1)-(2) = $\sum \dot{y}^2 = 499.36$	(1)-(2) = $\sum \dot{x}\dot{y} = 221.36$

$$r = \frac{\sum \dot{x}\dot{y}}{\sqrt{\sum \dot{x}^2 \sum \dot{y}^2}} = \frac{221.36}{\sqrt{386.40 \times 499.36}} = 0.5039$$

The steps taken in the calculation of Table 13.2.3.1 are:

1. Addition of the frequencies in each column (f_x) and each row (f_y).
2. Construction of a convenient coding of the number of hooks owned (x') and gillnet units owned (y').
3. Calculation of the marginal magnitudes $\sum x'f_x$, $\sum y'f_y$ and $\sum x'y'f_{xy}$.

In table 13.2.3.2 the value of r is calculated.

13.3 Linear regression

When we examine the data provided by a sample from a bivariate population one of the things we wish to ascertain first is whether there is any evidence of association between the variates of samples. If there is no association between the variates they would be called statistically independent. If the variates do tend to vary together in a rough sort of way they would be called stochastically or statistically related. If they are statistically related the next thing we want to find out is the type of

relationship between them. If a series of points (x_i, y_i) (population observations) lie approximately in a straight line it is reasonable to express the relationship between them as,

$$y_i = \alpha+\beta x_i+u_i$$

The problem in the present case consists of estimating the parameters α, β, by using the data of randomly selected samples from the parent populations. It should be noted that the line of regression of y on x gives an estimate of the average value of a dependent variable y associated with a value of the independent variable x. The regression line is a line of best fit to the empirical data when values of x are accurately known. Specifically, the problem consists of estimating the parameters α and β, taking into account all the observations, which will yield a line passing as close as possible to all of them.

The classical solution to the above problem was proposed by Gauss and is known as the method of least squares[1/]. Values of the coefficients are calculated in such a way that the sum of the squared deviations of the observed values from the fitted line ($\sum_i u_i^2$) is a minimum for the given set of observations. These values are easily found by application of the differential calculus, because $\sum_i u_i^2$ is at a minimum when,

$$\frac{\partial}{\partial\alpha}(\sum_i u_i^2) = 0 = \frac{\partial}{\partial\beta}(\sum_i u_i^2)$$

This procedure yields a system of two normal equations from which the estimated values of α and β are derived.

13.3.1 The regression line of y on x

A regression line fitted to the empirical data by the method of least squares is usually considered by statisticians to be the best with which to estimate the values of the dependent variable (y) when the values of the independent variable (x) are known (see Figure 13.3.1.1).

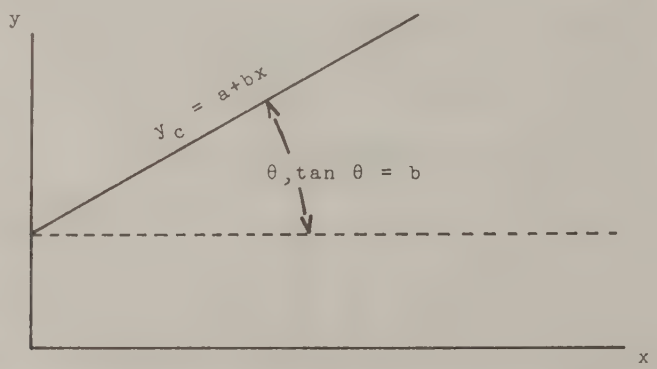


Figure 13.3.1.1 y_c as a linear function of x (b is its slope)

The system of normal equations for the calculation of a and b is given by,

$$\sum_i y_i = na+b\sum_i x_i$$

$$\sum_i y_i x_i = a\sum_i x_i+b\sum_i x_i^2$$

1/ The method of least squares is based on a number of assumptions regarding both the error term and the variables i.e. the variables x, y are free of measurement errors, the errors are independent among them and the errors are normally distributed.

By solving the above system of normal equations we get,

$$a = \bar{y} - b\bar{x}$$

and,

$$b = \frac{\begin{vmatrix} n & \sum_i y_i \\ \sum_i x_i & \sum_i y_i x_i \end{vmatrix}}{\begin{vmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{vmatrix}} = \frac{n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{n\sum_i x_i^2 - (\sum_i x_i)^2}$$

The coefficient of regression of y on x shows the average increase of the dependent variable for each additional unit of the independent variable.

Algebraical transformations:

$$b = \frac{\sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n}}{\left(\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}\right)}$$

or,

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum_i \dot{x}_i \dot{y}_i}{\sum_i \dot{x}_i^2}$$

or,

$$b = r\frac{s_y}{s_x}$$

Example

The table below gives the fish catch of 10 FEU's on a given day (x : kg) and the quantity of the catch which entered the commercial channel (y : kg). Estimate the linear regression equation of y on x and the value of the linear correlation coefficient.

- kg -

FEU	x	y	Remarks
1	25	20	($\sum x_i y_i$ = 17463)
2	30	22	
3	29	21	
4	40	35	
5	38	33	($\sum x_i^2$ = 20430)
6	44	40	
7	46	40	
8	52	45	
9	58	50	($\sum y_i^2$ = 14964)
10	70	60	
	432	366	

1. Linear regression equation,

$$y_c = a+bx$$

i) $$b = \frac{n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{n\sum_i x_i^2 - (\sum_i x_i)^2} = \frac{10\times 17463-(432)(366)}{10\times 20430-(432)^2} = 0.93$$

ii) $$a = 36.6-0.93\times 43.2 = -3.58$$

and,

$$y_c = -3.58+0.93x$$

2. Linear correlation coefficient,

$$r = \frac{n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{\sqrt{\{n\sum_i x_i^2-(\sum_i x_i)^2\}\{n\sum_i y_i^2-(\sum_i y_i)^2\}}}$$

$$= \frac{10\times 17463-(432)\times(366)}{\sqrt{\{10\times 20430-(432)^2\}\{10\times 14964-(366)^2\}}}$$

$$= 0.99$$

13.3.2 The mean square error

The mean square error provides a measure of the divergence of the actual values of the dependent variable from their estimated or computed values. This measure is analogous to a standard deviation and gives an idea, in absolute terms, of the dependability of estimates. The estimated mean square error (sample data) is given by,

$$s_{y\cdot x} = +\sqrt{s^2_{y\cdot x}}$$

where,

$$s^2_{y\cdot x} = \frac{1}{n-2}\sum_i (y_i-y_c)^2$$

or,

$$s^2_{y\cdot x} = \frac{1}{n-2}(\sum_i y_i^2 - a\sum_i y_i - b\sum_i x_i y_i)$$

or,

$$s^2_{y\cdot x} = \frac{1}{n-2}\{\sum_i y_i^2 - (a\sum_i y_i + b\sum_i x_i y_i)\}$$

Example

By using the data of the above example (section 13.3.1) calculate the value of $s_{y\cdot x}$.

$$s_{y\cdot x} = +\sqrt{s^2_{y\cdot x}}$$

where,

$$s^2_{y\cdot x} = \frac{1}{8}\left(14964-\{(-3.58)\times 366+0.93\times 17463\}\right) = 4.21$$

and,

$$s_{y\cdot x} = +\sqrt{4.21} = 2.05$$

13.3.3 The standard error of regression coefficient

The estimated standard error of b is given by,

$$\hat{\sigma}_b = +\sqrt{\hat{\sigma}_b^2}$$

where,

$$\hat{\sigma}_b^2 = \frac{s_{y\cdot x}^2}{\sum(x_i-\bar{x})^2} = \frac{s_{y\cdot x}^2}{(n-1)s_x^2}$$

The estimated confidence limits for the population regression coefficient (β), for a given probability level, is given by,

$$b-t_{\nu\cdot a}\hat{\sigma}_b<\beta<b+t_{\nu\cdot a}\hat{\sigma}_b \quad , \qquad (\nu = n-2)$$

Also, the estimated parallel magnitudes b, $\hat{\sigma}_b$ can be used to test the hypothesis that Ho : $\beta = \beta_0$.

Example

The calculated value of b based on a sample of 10 observations (x_i, y_i) equals 0.93. i) Test the hypothesis that Ho : $\beta = 0$. ii) If $\beta \neq 0$, estimate the confidence limits for β ($\alpha = 0.05$) (Note: $s_{y\cdot x}^2 = 0.16$, $\sum_i x_i = 200$, $\sum_i x_i^2 = 4382$).

Testing the hypothesis Ho : $\beta = 0$

(1) $$\hat{\sigma}_b^2 = \frac{s_{y\cdot x}^2}{\{\sum x_i^2 - \frac{(\sum x_i)^2}{n}\}} = \frac{0.16}{\{4382-\frac{(200)^2}{10}\}} = 0.000419$$

$$\hat{\sigma}_b = \sqrt{0.000419} = 0.0205$$

(2) $$t = \frac{|0.93-0|}{0.0205} = 45.36$$

$$t_{\nu=8,\ a=0.05} = 2.31 \ , \quad t_{\nu=8,\ a=0.01} = 3.36$$

Conclusion: The hypothesis that $\beta = 0$ is rejected.

Estimated confidence limits for β, ($a = 0.05$)

$0.93-2.31\times0.0205<\beta<0.93+2.31\times0.0205$

$0.8826<\beta<0.9774$

13.4 Further linear correlation and regression theory

13.4.1 The correlation coefficient and explained variation

Assume that by using the sample observations $\begin{matrix} x_1, x_2, \cdots, x_n \\ y_1, y_2, \cdots, y_n \end{matrix}$ a linear regression line was fitted to the empirical data (see figure following on next page).

One measure of variability, called total variation, is the sum of squares of the deviations of y values from their mean,

$$\sum_i (y_i-\bar{y})^2$$

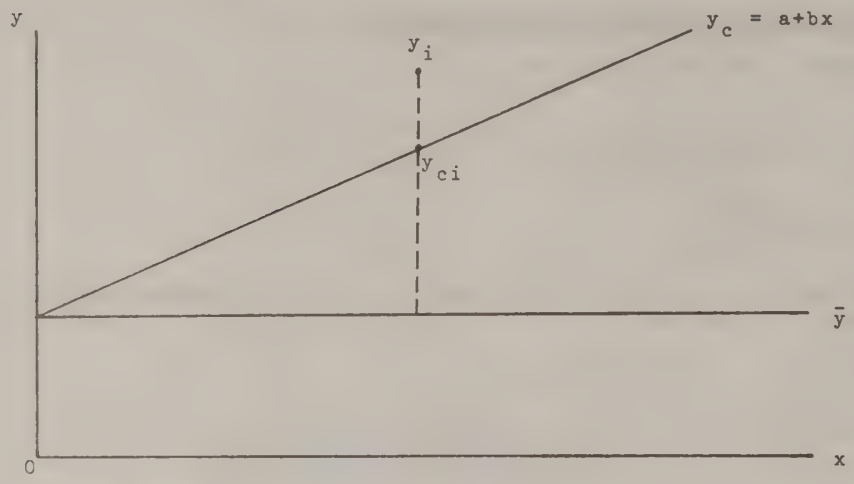


Figure 13.4.1.1

This total variation can be broken up into two parts: (1) that which we have failed to explain by the regression line, and (2) that which we have explained by the regression line. Specifically,

$$y_i-\bar{y} = (y_i-y_{ci})+(y_{ci}-\bar{y}) = d_{y\cdot x}+(y_{ci}-\bar{y})$$

and,

$$\sum_i (y_i-\bar{y})^2 = \sum_i d^2_{y\cdot x}+\sum_i (y_{ci}-\bar{y})^2$$

$$\left(\begin{matrix}\text{Total}\\ \text{variation}\end{matrix}\right) = \left(\begin{matrix}\text{Unexplained}\\ \text{variation}\end{matrix}\right)+\left(\begin{matrix}\text{Explained}\\ \text{variation}\end{matrix}\right)$$

Dividing the terms of both sides of the above equation by $\sum_i (y_i-\bar{y})^2$, we express the terms of the second part of the equation as ratios to one,

$$\frac{\sum_i (y_i-\bar{y})^2}{\sum_i (y_i-\bar{y})^2} = \frac{\sum_i d^2_{y\cdot x}}{\sum_i (y_i-\bar{y})^2}+\frac{\sum_i (y_{ci}-\bar{y})^2}{\sum_i (y_i-\bar{y})^2}$$

$$1 = (1-r^2)+r^2$$

The magnitude r^2 is the coefficient of determination. The coefficient of correlation r is the square root of the coefficient of determination.

Algebraical transformations:

We have,

$$1 = (1-r^2)+r^2$$

and,

$$r^2 = 1-(1-r^2)$$

or,

$$r^2 = 1-\frac{\sum_i d^2_{y\cdot x}}{\sum_i (y_i-\bar{y})^2}$$

or,

$$r^2 = 1-\frac{\{\sum_i y_i^2-(a\sum_i y_i+b\sum_i x_i y_i)\}}{\{\sum_i y_i^2-\frac{(\sum_i y_i)^2}{n}\}}$$

or,

$$r^2 = 1-\frac{(n-2)s_{y\cdot x}^2}{(n-1)s_y^2}$$

Example

The table below gives the total number of fishing operations (x) of 10 FEU's within a quarter and the obtained total fish catch (y : kg). Estimate the linear regression equation and the value of the coefficient of determination(r^2).

FEU's	x	y	Remarks
1	61	610	($\sum_i x_i y_i$ = 592860)
2	70	580	
3	80	700	
4	75	800	
5	79	750	($\sum_i x_i^2$ = 60686)
6	80	830	
7	82	800	
8	81	750	($\sum_i y_i^2$ = 5824900)
9	83	850	
10	85	900	
	776	7570	

1. $y_c = a+bx$

$$b = \frac{n\sum_i x_i y_i-(\sum_i x_i)(\sum_i y_i)}{n\sum_i x_i^2-(\sum_i x_i)^2} = \frac{10\times592860-776\times7570}{10\times60686-(776)^2}$$

$$= 11.59$$

$$a = 757-11.59\times77.6 = -142.38$$

$$y_c = -142.38+11.59x$$

2. i) $$r^2 = 1-\frac{\{\sum_i y_i^2-(a\sum_i y_i+b\sum_i x_i y_i)\}}{\{\sum_i y_i^2-\frac{(\sum_i y_i)^2}{n}\}}$$

$$= 1-\frac{\{5824900-\{(-142.38)\times7570+11.59\times592860)\}\}}{\{5824900-\frac{(7570)^2}{10}\}}$$

$$= 1-\frac{31469.20}{94410.00} = 0.667$$

ii) Variation:

$$1 = 0.333 + 0.667$$

$$\left(\begin{matrix}\text{Total}\\\text{variation}\end{matrix}\right) = \left(\begin{matrix}\text{Unexplained}\\\text{variation}\end{matrix}\right)+\left(\begin{matrix}\text{Explained}\\\text{variation}\end{matrix}\right)$$

iii) $r = \sqrt{r^2} = \sqrt{0.667} = 0.817$

13.4.2 Fiducial limits of y_c for given x_0

The fiducial limits of y_c for a given x_0 and given probability level, are,

$$y_{c/x_0} - t_\nu \cdot a^{s}\hat{R} < Y_{c/x_0} < y_{c/x_0} + t_\nu \cdot a^{s}\hat{R}$$

where,

$$s_{\hat{R}} = +\sqrt{s^2_{\hat{R}}}$$

and,

$$s^2_{\hat{R}} = s^2_{y\cdot x}\left\{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\{\sum_i x_i^2 - \frac{(\sum x_i)^2}{n}\}}\right\}$$

Example

by using available sample data the linear regression equation of y on x was estimated. Estimate the fiducial limits of y_c for x = 40 ($\bar{x}$ = 20, $\sum_i \dot{x}^2$ = 382).

1. Estimated regression equation

   b = 0.93 a = -0.1 n = 10

   and,

   $y_c = -0.1+0.93x$

2. i) $s^2_{y\cdot x} = 0.16$

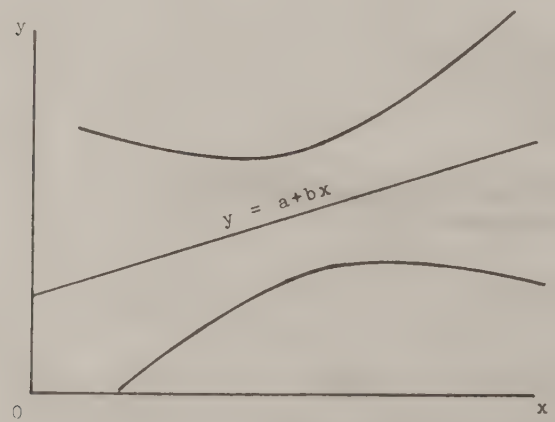
   $$s_{\hat{R}} = \sqrt{0.16\{\frac{1}{10} + \frac{(40-20)^2}{382}\}} = 0.43$$

   ii) $y_{c/x=40} = -0.1+0.93\times 40 = 37.10$

   iii) $37.1-2.31\times 0.43 < Y_{c/x=40} < 37.1+2.31\times 0.43$

   $36.11 < Y_{c/x=40} < 38.09$

The figure below gives an example of the fiducial belt for a linear regression.



13.4.3 Correlation ratio

Where a single qualitative factor is to be considered as an independent variable, and another quantitative factor as the dependent, the regression relation may be determined by grouping the observations according to the category of the qualitative factor and calculating the average value of the dependent factor for each group.

The intensity of correlation between the quantitative dependent factor and the qualitative independent factor is measured by the correlation ratio $\eta_{y \cdot x}$:

$$\eta_{y \cdot x} = +\sqrt{\eta^2_{y \cdot x}}$$

where,

$$\eta^2_{y \cdot x} = \frac{\sum_h n_h (\bar{y}_h - \bar{\bar{y}})^2}{\sum_{hi} \sum (y_{hi} - \bar{\bar{y}})^2}$$

In the above equation suffix h stands for the h^{th} group of observations and suffix i for the i^{th} observation. Also,

$$\bar{y}_h = \frac{1}{n_{hi}} \sum y_{hi}$$

$$\bar{\bar{y}} = \frac{1}{n_{hi}} \sum \sum y_{hi} \quad , \quad (n = \sum_h n_h)$$

Example

The table below gives the recorded surface temperature (y : C^o) of a given day at the sampling stations within the established three limnological zones (x : LZ) at lake L. Calculate (estimate) the value of $\eta_{y \cdot x}$.

- hypothetical data -

Sampling stations	Limnological zones x: L-1	L-2	L-3	Remarks
i=1	28	27	26	$(T = \sum_h T_h = 523)$
i=2	29	26	28	
i=3	30	28	28	
i=4 y: →	30	30	30	$(\sum_h \frac{T_h^2}{n_h} = 15200.17)$
i=5	30	29	30	
i=6	31	31	32	
T_h	178	171	174	$(\sum_{hi} \sum y^2_{hi} = 15245)$
T_h^2	31684	29241	30276	

Algebraical transformations:

$$\eta^2_{y \cdot x} = \frac{\sum_h n_h (\bar{y}_h - \bar{\bar{y}})^2}{\sum_{hi} \sum (y_{hi} - \bar{\bar{y}})^2}$$

since,

$$\sum_h n_h (\bar{y}_h - \bar{\bar{y}})^2 = \sum_h \frac{T_h^2}{n_h} - \frac{T^2}{n}$$

and,

$$\sum_{hi} \sum (y_{hi} - \bar{\bar{y}})^2 = \sum_{hi} \sum y^2_{hi} - \frac{(\sum_{hi} \sum y_{hi})^2}{n}$$

Substituting in the original formula we get,

$$\eta^2_{y\cdot x} = \{\sum_h \frac{T_h^2}{n_h} - \frac{T^2}{n}\} \Big/ \{\sum_{hi} y^2_{hi} - \frac{(\sum\sum y_{hi})^2}{n}\}$$

and,

$$\eta^2_{y\cdot x} = \frac{15200-15196}{15245-15196} = 0.0816$$

and,

$$\eta_{y\cdot x} = 0.2857$$

The estimated correlation ratio 0.2857 indicates some relation between the limnological zones of the lake and the surface temperature, but a rather low one, with the relation explaining only eight percent ($=\eta^2_{y\cdot x}$) at best of the variance in surface temperature between the limnological zones.

13.4.4 The linear regression of x on y

The estimated regression equation of x on y is given by,

$$x_c = a' + b'y$$

where,

$$a' = \bar{x} - b'\bar{y}$$

$$b' = \frac{n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{\{n\sum_i y_i^2 - (\sum_i y_i)^2\}}$$

etc.

It should be noted that the value of the correlation coefficient can be computed from the slopes of the two lines according to the equation,

$$r = \sqrt{bb'}$$

It follows from this equation that when r = 1, b = 1/b' and therefore the two regression lines will coincide. This property of the two lines can be used to estimate graphically the closeness of correlation.

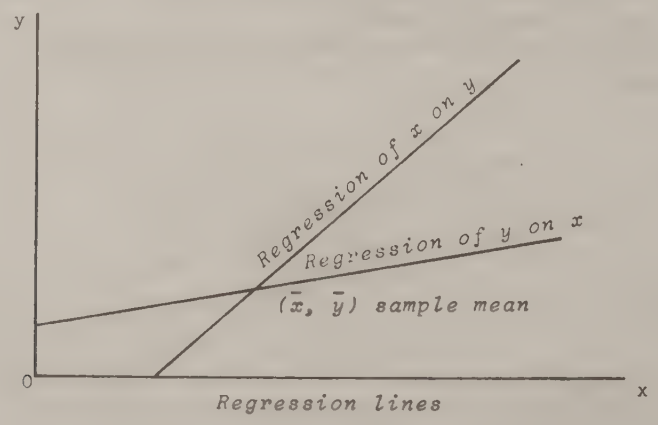


Figure 13.4.4.1

13.5 Linear multiple regression

13.5.1 Introduction

It frequently happens that a variable for which we wish to write a regression depends not on a single independent variable but on two or more independent variables. Or, looking at it from another angle, using more than one independent variable may enable us to predict the dependent variable with more accuracy than could be done from a single independent variable. This requires the use of multiple regression techniques.

13.5.2 Partial correlation and multivariate regression

When we are dealing with more than two variables the names of partial correlation (instead of correlation) and multivariate regression (instead of regression) are used. Specifically, in the problem of partial correlation we are interested in assessing the interdependence of two of the variates, after eliminating the influence of all the others. In multivariate regression we examine the influence on one of the variates by the others in the set.

13.5.3 The type of equation (linear multiple regression)[1/]

The first step in the correlation procedure is to obtain an equation which includes the surveyed variates. If we are dealing with three variables the equation type[2/] is,

$$Y_c = a + b_1 X_1 + b_2 X_2$$

If the three variables are measured from their means the above equation becomes,

$$\dot{y}_c = b_1 \dot{x}_1 + b_2 \dot{x}_2$$

The coefficients b_1 and b_2 are called partial regression coefficients. A partial regression coefficient indicates the effect on Y_c of a change in the accompanying independent variable when allowance has been made for the other independent variable(s).

13.5.4 The system of normal equations

When the variates are measured from their means the system of normal equations is,

	(3) : (1) (2)
(1)	$\sum_i \dot{y}_i \dot{x}_{1i} = (\sum_i \dot{x}_{1i}^2) b_1 + (\sum_i \dot{x}_{1i} \dot{x}_{2i}) b_2$
(2)	$\sum_i \dot{y}_i \dot{x}_{2i} = (\sum_i \dot{x}_{1i} \dot{x}_{2i}) b_1 + (\sum_i \dot{x}_{2i}^2) b_2$

By solving the system of the above equations the values of b_1 and b_2 are calculated.

1/ A linear multiple regression model can be applied to empirical data when the relationship between the dependent and each of the independent variables is linear.

2/ Another notation used for the equation is,

$$X_{1.23} = a_{1.23} + b_{12.3} X_2 + b_{13.2} X_3$$

where suffixes 1, 2, 3 stand for the three survey variates.

Example[1/]

The table below gives the data concerning the relationship between total fish catch (Y : kg/day) of 10 FEU's at a given lake and of the number of hours fished (X_1) and the number of hooks used (X_2). Estimate the linear multiple regression equation of Y on X_1 and X_2.

- hypothetical data -

FEU's	Y_i (kg)	X_{1i} (hr)	X_{2i} (No. of hooks)	$\dot{y}_i = Y_i - \bar{Y}$	$\dot{x}_{1i} = X_{1i} - \bar{X}_1$	$\dot{x}_{2i} = X_{2i} - \bar{X}$	$\dot{y}_i\dot{x}_{1i}$	$\dot{y}_i\dot{x}_{2i}$	$\dot{x}_{1i}\dot{x}_{2i}$	$\dot{x}^2_{1i}$	$\dot{x}^2_{2i}$	$\dot{y}_i$
1	4	2	32	-1	-2	-11	2	11	22	4	121	-1
2	2	5	18	-3	1	-25	-3	75	-25	1	625	-3
3	7	3	52	2	-1	9	-2	18	-9	1	81	2
4	1	3	6	-4	-1	-37	4	148	37	1	1369	-4
5	4	4	42	-1	0	-1	0	1	0	0	1	-1
6	5	4	48	0	0	5	0	0	0	0	25	0
7	9	6	80	4	2	37	8	148	74	4	1369	4
8	5	2	32	0	-2	-11	0	0	22	4	121	0
9	6	4	52	1	0	9	0	9	0	0	81	1
10	7	7	68	2	3	25	6	50	75	9	625	2
Total	50	40	430				15	460	196	24	4418	$(\sum \dot{y}_i^2 = 52)$
Mean	5	4	43									

1. System of normal equations,

$$15 = 24b_1 + 196b_2$$

$$460 = 196b_1 + 4418b_2$$

2. Estimated partial regression coefficients,

$$b_2 = \frac{\begin{vmatrix} 24 & 15 \\ 196 & 460 \end{vmatrix}}{\begin{vmatrix} 24 & 196 \\ 196 & 4418 \end{vmatrix}} = \frac{24\times460 - 15\times196}{24\times4418 - 196\times196} = 0.1198$$

$$b_1 = -0.35$$

3. Multiple regression equation,

$$(Y_c - 5) = -0.35(X_{1i} - 4) + 0.1198(X_{2i} - 43)$$

and,

$$Y_c = 1.249 - 0.35X_{1i} + 0.1198X_{2i}$$

1/ See also, G.P. Bazigos, Efficiency of Different Sampling Methods for Large Scale Biological and Fishery Statistical Sample Surveys at Large African Lakes: 1. Cove-Rotenone Sample Survey at Kariba Lake, St.E./1, UNDP/SF/ZAM/11, April 1972. G.P. Bazigos, The Yield Pattern at Lake Nasser, St.S./1, UNDP/SF/EGY/66/558, September 1972.

4. The calculated value of $b_1 = -0.35$ indicates the normal change in fish catch associated with a unit change in fishing time (one hour). The calculated value of $b_2 = 0.1198$ indicates the normal change in fish catch associated with the use of an additional hook in fishing operations.

13.5.5 Gauss multipliers

If all we want is a regression equation the problem is completed. This, however, is seldom the case. Most of the time we will want to make tests of significance and provide a measure of the precision of the regression. For this reason it is desirable to calculate Gauss multipliers or c-multipliers.

Computation of the c-multipliers involves the solution of simultaneous equations. The equations are very similar to the equations used to determine the partial regression coefficients. The sum of squares and crossproducts are arranged in exactly the same order except that the crossproducts involving Y are omitted. Thus, the basic set of equations for a regression problem with two independent variables would be,

	First set		Second set
$\sum_i \dot{x}_{1i}^2 + \sum_i \dot{x}_{1i}\dot{x}_{2i}$	= 1	: =	0
$\sum_i \dot{x}_{1i}\dot{x}_{2i} + \sum_i \dot{x}_{2i}^2$	= 0	: =	1

There would be two sets of these equations (with three independent variables there would be three sets of three equations each, with four independent variables there would be four sets of equations each, etc.).

The number of multipliers needed is equal to the number of sums of squares and crossproducts involving the independent variables only. For our example the multipliers are,

$$c_{11},\ c_{12},\ c_{22},\ (c_{12} = c_{21})$$

The suffixes in the multipliers stand, the first for the set of equations and the second for the X-column (see system of normal equations),

	(1)	(2)	
First set:	$c_{11}(\sum_i \dot{x}_{1i}^2)$	$+c_{12}(\sum_i \dot{x}_{1i}\dot{x}_{2i})$	= 1
	$c_{11}(\sum_i \dot{x}_{1i}\dot{x}_{2i})+c_{12}(\sum_i \dot{x}_{2i}^2)$		= 0

Second set:	$c_{21}(\sum_i \dot{x}_{1i}^2)$	$+c_{22}(\sum_i \dot{x}_{1i}\dot{x}_{2i})$	= 0
	$c_{21}(\sum_i \dot{x}_{1i}\dot{x}_{2i})+c_{22}(\sum_i \dot{x}_{2i}^2)$		= 1

By substituting the appropriate sums of squares and crossproducts of our example we get,

First set:		Second set:	
$24c_{11} + 196c_{12}$	= 1	$24c_{21} + 196c_{22}$	= 0
$196c_{11} + 4418c_{12}$	= 0	$196c_{21} + 4418c_{22}$	= 1

Solving each set simultaneously we get,

$$c_{11} = 0.065346$$

$$c_{12} = -0.002899 = c_{21}$$

$$c_{22} = 0.000355$$

By using the c-multipliers we can calculate the partial regression coefficients, the formulae[1] are,

$$b_1 = c_{11}(\sum_i \dot{x}_{1i}\dot{y}_i) + c_{12}(\sum_i \dot{x}_{2i}\dot{y}_i)$$

$$b_2 = c_{21}(\sum_i \dot{x}_{1i}\dot{y}_i) + c_{22}(\sum_i \dot{x}_{2i}\dot{y}_i)$$

Substituting the respective values in the above equations we get,

$$b_1 = 0.065346 \times 15 - 0.002899 \times 460 = -0.35$$

$$b_2 = -0.002899 \times 15 + 0.000355 \times 460 = 0.1198$$

13.5.6 Mean square error

After working out the multiple regression equation by which values of the dependent variable can be estimated from those for the two independent variables, it is desirable to estimate the value of the mean square error ($s_{y \cdot 12}$) to compare how closely such estimates agree with the empirical values,

$$s_{y \cdot 12} = +\sqrt{s^2_{y \cdot 12}}$$

and,

$$s^2_{y \cdot 12} = \frac{1}{n-k-1}\{\sum_i \dot{y}_i^2 - (b_1 \sum_i \dot{y}_i \dot{x}_{1i} + b_2 \sum_i \dot{y}_i \dot{x}_{2i})\}$$

where,

n : number of sets of observations in the sample.

k : number of independent variables.

Example

By using the data of our example (section 13.5.3) calculate the value of $s_{y \cdot 12}$,

$$s^2_{y \cdot 12} = \frac{1}{10-3}\{52 - (-0.35 \times 15 + 0.1198 \times 460)\}$$

$$= 0.306$$

$$s_{y \cdot 12} = +\sqrt{0.306} = 0.55$$

13.5.7 Analysis of total variation into components

The total variation of the individual values of Y about the mean value of Y($\bar{Y}$), is given by,

$$q_3 = \sum_i (Y_i - \bar{Y})^2 = \sum_i \dot{y}_i^2 \qquad (d.f. = n-1)$$

1/ Note that in the formulae each equation uses all of the crossproducts of $\dot{y}$ with each independent variable. The c-multiplier by which each crossproduct is multiplied is identified by its suffixes. The first suffix stands for the desired regression coefficient, the second is the same as the $\dot{x}$ suffix in the crossproduct.

The variation in Y, that is explained by the regression of Y on X_1 and X_2, is symbolized by q_1 and is calculated by the formula,

$$q_1 = b_1 \sum_i \dot{y}_i \dot{x}_{1i} + b_2 \sum_i \dot{y}_i \dot{x}_{2i} \qquad (d.f. = k)$$

The variation in Y that is still left unexplained even after fitting the regression is equal to,

$$q_2 = q_3 - q_1 \qquad (d.f. = n-k-1)$$

The table below gives the analysis of total variation of our example (section 13.5.3) into components.

Source of variation	d.f.	Sum of squares	Mean squares	F-values		
				Estimated	Tabulated 5(%)	Tabulated 1(%)
Due to regression	2	$q_1 = 49.86$	24.93	80.4194**	4.74	9.55
Residual (i.e. unexplained)	7	$q_2 = 2.14$	0.31		(Highly-significant)	
Total	9	$q_3 = 52.00$	-			

13.5.8 Estimated confidence limits for the population partial regression coefficients

By using the sample observations estimates can be calculated for the confidence limits of the population partial regression coefficients (β_i),

$$b_i - t_{\nu \cdot a} s_{bi} < \beta_i < b_i + t_{\nu \cdot a} s_{bi} \qquad (\nu = n-k-1)$$

where,

$$s_{bi} = s_{y \ldots} \sqrt{c_{ii}}$$

In our case (example of section 13.5.3) the estimated confidence limits for β_1, β_2 are respectively (P = 95 percent),

$$-0.35 - 2.36 \times 0.55 \sqrt{0.065346} < \beta_1 < -0.35 + 2.36 \times 0.55 \sqrt{0.065346}$$

$$-0.68 < \beta_1 < -0.02$$

$$0.1198 - 2.36 \times 0.55 \sqrt{0.000355} < \beta_2 < 0.1198 + 2.36 \times 0.55 \sqrt{0.000355}$$

$$0.095 < \beta_2 < 0.144$$

Also, the sample observations can be used to test the hypothesis that Ho : $\beta_i = \beta_0$.

$$t = \frac{|b_i - \beta_0|}{s_{bi}}$$

etc. (see also footnote on page 113).

13.5.9 The significance of the individual variables

A constant problem in multiple regression analysis is the question of whether or not a particular independent variable makes a significant contribution. For example, in our case we may want to know whether or not X_2 made a significant contribution to the regression. This can be done with c-multipliers:

1. We calculate what the sum of squares attributable to regression (q_1) would be if X_2 were deleted. The formula for the adjusted sum of squares (q_{1a}) is,

$$q_{1a} = q_1 - \frac{b_2^2}{c_{22}}$$

2. After the deletion of X_2 from the regression, the residual sum of squares (adjusted residual) would be,

   $q_{2a} = q_3 - q_{1a}$

   The reduction was made at a cost of one degree of freedom.

3. We estimate the respective F-value and compare it with the tabulated value (see table below).

Source of variation	d.f.	Sum of squares	Mean of squares	F-values		
				Estimated	Tabulated 5(%)	Tabulated 1(%)
Adjusted residual	8	$q_{2a} = 52.00-9.43 = 42.57$				
Unadjusted residual	7	$q_2 = 2.14$	0.31	** 135	5.59	12.25
Reduction due to X_2	1	$q = q_{2a}-q_2 = 40.43$	40.43		(Highly-significant)	

The contribution of X_1 can be tested in a similar manner. The formula for the adjusted sum of squares with X_1 deleted would be,

$$q_{1a} = q_1 - \frac{b_1^2}{c_{11}}$$

etc.

13.5.10 The multiple correlation coefficient

As we have discussed, the coefficient of determination offers a measure of the variation in the dependent variable which can be explained by variation in the independent variable(s). In the case of multiple linear regression the formula for estimating the multiple coefficient of determination is,

$$\hat{R}^2_{y\ldots} = 1 - \frac{(n-k-1)s^2_{y\ldots}}{(n-1)s^2_y}$$

In our example the estimated value of $\hat{R}^2_{1.23}$ is,

$$\hat{R}^2_{y\cdot 12} = 1 - \frac{(n-k-1)s^2_{y\cdot 12}}{(n-1)s^2_y}$$

or,

$$\hat{R}^2_{y\cdot 12} = \frac{b_1 \sum_i \dot{y}_i \dot{x}_{1i} + b_2 \sum_i \dot{y}_i \dot{x}_{2i}}{\sum_i \dot{y}_i^2}$$

$$= \frac{-0.35 \times 15 + 0.1198 \times 460}{52} = 0.9588$$

The multiple correlation coefficient is the square root of the multiple coefficient of determination. In our example,

$$\hat{R}_{y\cdot 12} = +\sqrt{\hat{R}^2_{y\cdot 12}} = \sqrt{0.9588} = 0.9792$$

The coefficient of multiple correlation measures the combined importance of the several independent variables as a means of explaining the differences in the dependent variable.

13.5.11 The reliability of the multiple correlation coefficient[1/]

The reliability of coefficients of multiple correlation varies not only with the correlation and size of sample, but also with the number of independent variables. Specifically, when dealing with multiple correlation coefficients, we are primarily interested in knowing whether a given $\hat{R}$ (or $\hat{R}^2$) is significant.

To ascertain whether a multiple coefficient of determination significantly exceeds zero the F-test is used,

$$F = \frac{\hat{R}^2_{y}}{(1-\hat{R}^2_{y\cdot\cdot\cdot})} \div \frac{k}{(n-k-1)} = \frac{\hat{R}^2_{y\cdot\cdot\cdot}}{1-\hat{R}^2_{y\cdot\cdot\cdot}} \times \frac{(n-k-1)}{k}$$

In our example (section 13.5.10) the estimated value of F is,

$$F = \frac{0.9588}{0.0412} \times \frac{7}{2} = \frac{6.7116}{0.0824} = 81.45$$

The tabulated F-value for $\nu_1 = 2$ and $\nu_2 = 7$ ($a = 0.05$) is 4.74. Since the estimated F-value is larger than the tabulated one we conclude that the estimated determination coefficient is highly significant at $a = 0.05$.

The significance of the multiple correlation coefficient can be tested by using the t-test,

$$t = \frac{\hat{R}_{y\cdot\cdot\cdot}}{\sqrt{1-\hat{R}^2_{y\cdot\cdot\cdot}}} \sqrt{\frac{n-k-1}{k}}$$

In our example the calculated value of t is,

$$t = \frac{0.9792}{\sqrt{0.0412}} \sqrt{\frac{7}{2}} = 9.0252$$

The tabulated t-value ($a = 0.05$, $\nu = n-k-1$) is $t_{\nu\cdot a} = 2.365$, highly significant.

13.5.12 Partial correlation coefficients

A coefficient of partial correlation measures the closeness of relationship between two variables of a multiple regression model when the other variates of the model have been held constant statistically. The sign of this coefficient of partial correlation is the same as the sign of the respective multiple regression coefficient (b_i). In statistical language, a partial correlation coefficient is referred to as first-order when one variable has been held constant, second-order when two variables have been held constant, etc. Simple coefficients are called zero-order coefficients, since no variables were held constant.

One of the various approaches used to estimate the partial regression coefficients is the method of zero-coefficients. In our example (section 13.5.3) the formulae for the partial correlation coefficients are,

---

1/ To test the significance between the estimated partial regression coefficients the t-test is used, in our example,

$$t = \frac{|b_1 - b_2|}{s_{b1-b2}}$$

where,

$$s_{b1-b2} = s_{y\cdot 12} \sqrt{c_{11}+c_{22}-2c_1c_2}$$

$$r_{y1.2} = \frac{r_{y1}-r_{y2}r_{12}}{\sqrt{1-r_{y2}^2}\sqrt{1-r_{12}^2}}$$

$$r_{y2.1} = \frac{r_{y2}-r_{y1}r_{12}}{\sqrt{1-r_{y1}^2}\sqrt{1-r_{12}^2}}$$

$$r_{12.y} = \frac{r_{12}-r_{y1}r_{y2}}{\sqrt{1-r_{y1}^2}\sqrt{1-r_{y2}^2}}$$

since,

$$r_{y1} = 0.425 \quad , \quad \sqrt{1-r_{y1}^2} = 0.905$$

$$r_{y2} = 0.960 \quad , \quad \sqrt{1-r_{y2}^2} = 0.280$$

$$r_{12} = 0.602 \quad , \quad \sqrt{1-r_{12}^2} = 0.798$$

we have,

$$r_{y1.2} = \frac{0.425-0.960\times0.602}{0.208\times0.798} = -0.921$$

$$r_{y2.1} = \frac{0.960-0.425\times0.602}{0.905\times0.798} = 0.975$$

$$r_{12.y} = \frac{0.602-0.425\times0.960}{0.905\times0.280} = 0.766$$

As one can see from the values of the estimated magnitudes the gross correlation between catches (Y) and fishing time (X_1) is $r_{y1} = 0.425$. Removing the effect of variations of X_2 (number of hooks) from both of them the relationship decreases materially to $r_{y1.2} = -0.921$, etc.

13.5.13 The reliability of partial correlation coefficients

For samples selected from normally distributed populations the significance of a partial correlation coefficient i.e. whether a partial correlation coefficient significantly exceeds zero, can be tested by using the t-test,

$$t = \frac{r...}{\sqrt{1-r^2...}}\sqrt{n-p-2} \quad , \quad (p = \text{order of } r...)$$

By using the data of our example (section 13.5.11) the calculated values of t are,

$$t_1 = \frac{|-0.921|}{\sqrt{1-0.921^2}}\sqrt{7} = 6.25510, \quad \text{significant}$$

$$t_2 = \frac{0.975}{\sqrt{1-0.975^2}}\sqrt{7} = 11.60939, \quad \text{significant}$$

$$t_3 = \frac{0.766}{\sqrt{1-0.766^2}}\sqrt{7} = 3.15264, \quad \text{significant}$$

The tabulated t-value is $t_{\nu=7,\ a=0.05} = 2.365$.

Also, the transformer z (see section 13.2.1) can be used both to estimate the confidence limits for a population partial correlation coefficient ($\rho...$) and to test the hypothesis that Ho : $\rho... = \rho...(o)$.

1. $z\ldots - t_{\nu a}\hat{\sigma}_{z\ldots} < Z < z\ldots + t_{\nu a}\hat{\sigma}_{z\ldots}$

and,

$$1^{r}\ldots < \rho\ldots < {}_{2}r\ldots$$

2. $$t = \frac{|z\ldots|}{\sigma_{z\ldots}}$$

where,

$$\hat{\sigma}_{z\ldots} = +\sqrt{\hat{\sigma}^2_{z\ldots}}$$

and,

$$\hat{\sigma}^2_{z\ldots} = \frac{1}{n-p-3}$$

etc.

13.6 Polynomial regression

In section 13.3 we have limited our discussion of regression to bivariate distributions where the regression curves were straight lines. If the regression is not linear we call it curvilinear. The simplest type of curvilinear equation is that in which one of the variates is a polynomial function of the other,

$$y_c = a + bx_i + cx_i^2 + \ldots + \lambda x_i^k$$

If the regression equation of y on x (or x on y) is of this form we have polynomial regression. Specifically, if k = 2 we have a second-degree curve (parabola) with the estimating equation,

$$y_c = a + bx_i + cx_i^2$$

and if k = 3 we have a third-degree curve with the estimating equation,

$$y_c = a + bx_i + cx_i^2 + dx_i^3$$

When we are dealing with curvilinear regression the coefficient of linear correlation r_{xy}, which is a measure of the extent to which any relationship between the variates tends towards linearity, is no longer a suitable measure of correlation.

For a second-degree curve, a system of three normal equations is required (method of least squares),

	(4) : (1) + (2) + (3)
(1)	$\sum_i y_i = na + b\sum_i x_i + c\sum_i x_i^2$
(2)	$\sum_i y_i x_i = a\sum_i x_i + b\sum_i x_i^2 + c\sum_i x_i^3$
(3)	$\sum_i y_i x_i^2 = a\sum_i x_i^2 + b\sum_i x_i^3 + c\sum_i x_i^4$

The estimated coefficient of correlation $\hat{R}_{yx}$ is the square root of the estimated coefficient of determination R^2_{yx}, i.e.,

$$\hat{R}_{yx} = +\sqrt{R^2_{yx}}$$

The coefficient of determination is, as before, the ratio of the explained variation to total variation,

$$\hat{R}^2_{yx} = 1-\frac{(n-k-1)s^2_{y.x}}{(n-1)s^2_y}$$

where,

$$s^2_{y.x} = \frac{1}{n-k-1}\{\sum y_i^2-(a\sum y_i+b\sum x_i y_i+c\sum x_i^2 y_i)\}$$

is the mean square error.

For a third-degree curve, a system of four normal equations is required (method of least squares),

	(5)	(1) +(2) +(3) +(4)
(1)	$\sum y_i$	$= na +b\sum x_i +c\sum x_i^2 +d\sum x_i^3$
(2)	$\sum y_i x_i$	$= a\sum x_i +b\sum x_i^2 +c\sum x_i^3 +d\sum x_i^4$
(3)	$\sum y_i x_i^2$	$= a\sum x_i^2 +b\sum x_i^3 +c\sum x_i^4 +d\sum x_i^5$
(4)	$\sum y_i x_i^3$	$= a\sum x_i^3 +b\sum x_i^4 +c\sum x_i^5 +d\sum x_i^6$

If $\dot{x}_i = x_i-\bar{x}$ the above system of normal equations reduces to,

$$\sum y_i = na +c\sum \dot{x}_i^2$$

$$\sum y_i \dot{x}_i = b\sum \dot{x}_i^2 +d\sum \dot{x}_i^4$$

$$\sum y_i \dot{x}_i^2 = a\sum \dot{x}_i^2 +c\sum \dot{x}_i^4$$

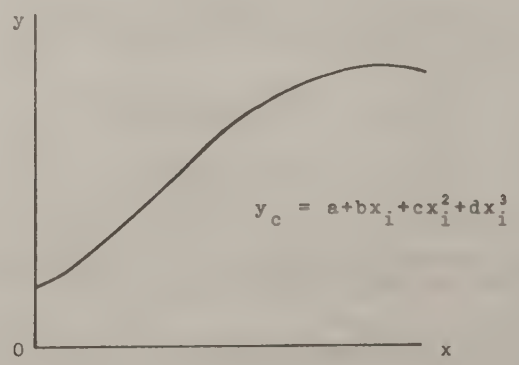
$$\sum y_i \dot{x}_i^3 = b\sum \dot{x}_i^4 +d\sum \dot{x}_i^6$$

The estimated coefficient of correlation $\hat{R}_{yx}$, is the square root of the estimated coefficient of determination $\hat{R}^2_{yx}$,

$$\hat{R}_{yx} = +\sqrt{\hat{R}^2_{yx}}$$

where,

$$\hat{R}^2_{yx} = 1-\frac{(n-k-1)s^2_{y.x}}{(n-1)s^2_y}$$



Example

The tabulated data (Table 13.6.1) gives the total number of gillnet units fished per FEU (day, do) and the average total fish catch (kg) per gillnet unit. Fit a second-degree curve to the empirical data and calculate the value of $\bar{R}_{xy}$ (data from a fishing camp at Volta Lake).

Table 13.6.1 Estimated number of gillnet units fished per fishing boat and average catches per fishing gillnet unit/day in a fishing camp at Volta Lake

Fishing boats	x_i 1/	y_i 2/	$\dot{x}_i = x_i - \bar{x}$	$\dot{x}_i y_i$	$\dot{x}_i^2$	$\dot{x}_i^4$	Remarks
1	5	1.20	-4.5	-5.40	20.25	410.0625	($\sum_i x_i = 95$, $\sum y_i = 18.10$)
2	6	1.60	-3.5	-5.60	12.25	150.0625	
3	7	1.90	-2.5	-4.75	6.25	39.0625	($\sum_i \dot{x}_i y_i = +0.56$)
4	8	2.10	-1.5	-3.15	2.25	5.0625	($\sum_i \dot{x}_i^2 y_i = 124.445$)
5	9	2.20	-0.5	-1.10	0.25	0.0625	
6	10	2.15	+0.5	+1.075	0.25	0.0625	($\sum_i \dot{x}_i^2 = 82.50$)
7	11	2.10	+1.5	+3.15	2.25	5.0625	
8	12	1.92	+2.5	+4.80	6.25	39.0625	($\sum_i \dot{x}_i^4 = 1208.6250$)
9	13	1.65	+3.5	+5.775	12.25	150.0625	($\sum_i y_i^2 = 33.9398$)
10	14	1.28	+4.5	+5.76	20.25	410.0625	
	95	18.10	0	+0.56	82.50	1208.6250	
	$\bar{x}$:9.5	$\bar{y}$:1.81					

1/ x : gillnet units fished per day

2/ y : fish catch (kg) per fished gillnet unit/day

System of normal equations,

(1) $\sum_i y_i = na + c\sum_i \dot{x}_i^2$

(2) $\sum_i y_i \dot{x}_i = b\sum_i \dot{x}_i^2$

(3) $\sum_i y_i \dot{x}_i^2 = a\sum_i \dot{x}_i^2 + c\sum_i \dot{x}_i^4$

and,

(1) $18.10 = 10a + 82.50c$

(2) $0.56 = 82.50b$

(3) $124.445 = 82.50a + 1208.6250c$

Estimating equations,

$b = 0.0068$

$$c = \frac{\begin{vmatrix} 10 & 18.10 \\ 82.50 & 124.445 \end{vmatrix}}{\begin{vmatrix} 10 & 82.50 \\ 82.50 & 1208.625 \end{vmatrix}} = \frac{1244.45-1493.25}{12086.25-6806.25} = -0.0471$$

$$a = 2.1984$$

$$y_c = 2.1984+0.0068(x_i-9.5) - 0.0471(x_i-9.5)^2$$

and,

$$y_c = -2.1170+0.9017x_i-0.0471x_i^2$$

The estimated coefficient of correlation is given by,

$$\hat{R}_{yx} = +\sqrt{\hat{R}^2_{yx}}$$

where,

$$\hat{R}^2_{yx} = 1-\frac{\{\sum_i y_i^2-(a\sum_i y_i+b\sum_i \dot{x}_i y_i+c\sum_i \dot{x}_i^2 y_i)\}}{\{\sum_i y_i^2-\frac{(\sum_i y_i)^2}{n}\}}$$

$$= 1-\frac{0.0063}{1.1788} = 0.9947$$

and,

$$\hat{R}_{yx} = 0.9973^{1/}$$

13.7 Regression of equations using transformations

It has been discussed that the discovery of a precise description of the relationship between two (or more) quantities is one of the problems of curve fitting. In some cases it is convenient to convert the readings for one or both variables into a different form. The most frequently used transformations involve logarithms, reciprocals, roots or powers and logarithms of logarithms. Frequently a transformation will show a linear relationship between the two converted series.

1/ If the sample size of the bivariate values is large, an approximation of the estimated variance of R_{yx} is given by,

$$s^2_{\hat{R}_{yx}} \simeq \frac{(1-\hat{R}^2_{yx})^2}{n-2}$$

The estimated confidence interval for R_{yx} at a given probability level is,

$$\hat{R}_{yx}-t_a s_{\hat{R}} < R < \hat{R}_{yx}+t_a s_{\hat{R}}$$

Also, the estimated parallel magnitudes $\hat{R}_{yx}$ and $s_{\hat{R}}$ can be used to test the hypothesis of significance for R_{yx},

$$t = \frac{|\hat{R}_{yx}-0|}{s_{\hat{R}}}$$

13.7.1 The log y, log x relationship[1/]

The estimating equation is of the type,

$$y_c = ax_i^b$$

or,

$$\log y_c = \log a + b(\log x_i)$$

The relationship portrays a straight line when both series are expressed in terms of logarithms.

The constants, log a and b, are obtained by solving simultaneously the normal equations,

	(3)	:	(1)	+	(2)
(1)	$\sum_i \log y_i$	=	$n\log a$		$+b\sum_i \log x_i$
(2)	$\sum_i (\log y_i)(\log x_i)$	=	$\log a\sum_i \log x_i$		$+b\sum_i (\log x_i)^2$

The estimated correlation coefficient is given by $\hat{R}_{\log y.\log x} = +\sqrt{\hat{R}^2_{\log x.\log y}}$

$$\hat{R}^2_{\log y.\log x} = 1 - \frac{(n-2)s^2_{\log y.\log x}}{(n-1)s^2_{\log y}}$$

where,

$$s^2_{\log y.\log x} = \frac{1}{n-2}\{\sum_i (\log y_i)^2 - (\log a\sum_i \log y_i) - b\sum_i (\log y_i)(\log x_i)\}$$

and,

$$s^2_{\log y} = \frac{1}{n-1}\{\sum_i (\log y_i)^2 - \frac{(\sum_i \log y_i)^2}{n}\}$$

Example

The table below (Table 13.7.1.1) gives the total number of hooks fished per FEU (x) at lake L (day, do) and the obtained fish catch (y : kg). Fit a log curve to the empirical data and estimate the value of the correlation coefficient.

1/ In fish biology, with any set of lengths and weights, the relationship is usually of the form,

$$w = al^b \qquad (w : \text{weight}, \quad l : \text{length})$$

or,

$$\log w = \log a + b(\log l)$$

When the exponent b = 3, the animal is growing without change in shape or specific gravity i.e. isometrically. Among fish an exponent of three is the exception, however, values from 2 to 3.5 are commonly found.

Table 13.7.1.1 Estimated number of hooks fished per FEU (x) at lake L (day, do) and obtained fish catch (y : kg)[1]

FEU's	x_i (no.)	y_i (kg)	log x_i	log y_i	Remarks
1	20	1.0	1.30103	0.00000	$\sum_i (\log y_i)(\log x_i) = 7.76535$
2	26	1.4	1.41497	0.14613	
3	34	2.0	1.53148	0.30103	
4	44	2.8	1.64345	0.44716	$\sum_i (\log x_i)^2 = 23.64470$
5	57	3.9	1.75587	0.59106	
6	74	5.5	1.86923	0.74036	
7	96	7.7	1.98227	0.88649	$\sum_i (\log y_i)^2 = 3.07977$
8	125	11.0	2.09691	1.04139	
			13.59521	4.15362	

Estimating the regression equation,

(3) : (1) + (2)	
(1)	$4.15362 = 8 \log a + 13.59521 b$
(2)	$7.76535 = 13.59521 \log a + 23.64470 b$

$\log a = -1.70075$ (or $\log a = \bar{2}.29925$)

$b = 1.30632$

$\log y_c = -1.70075 + 1.30632 \log x_i$

or,

$$y_c = 0.01992 x_i^{1.30632}$$

The estimated value of the coefficient of determination is,

$$\hat{R}^2_{\log y \cdot \log x} = 1 - \frac{0.00001}{0.92320} = 0.99999$$

The estimated value of the coefficient of correlation is,

$$\hat{R}_{\log y \cdot \log x} \simeq 1$$

13.7.2 The log y, x relationship

When correlating logarithms of y with x values, the estimating equation is of the type[2],

---

1/ Note that the model has a good fit to the empirical data when the values of x and y follow geometric progressions, i.e.,

$$\frac{x_2}{x_1} = \frac{x_3}{x_2} = \ldots = \frac{x_8}{x_7} = 1.30$$

and,

$$\frac{y_2}{y_1} = \frac{y_3}{y_2} = \ldots = \frac{y_8}{y_7} = 1.40$$

2/ The model has a good fit to the empirical data when the values of y follow a geometric progression and the values of x an arithmetic progression.

$$\log y = \log a + (\log b)x_i$$

or,

$$y = ab^{x_i}$$

The normal equations are,

	(3)	:	(1)	+	(2)
(1)	$\sum_i \log y_i$	=	$n(\log a)$	+	$(\log b)\sum_i x_i$
(2)	$\sum_i (\log y_i)x_i$	=	$\log a \sum_i x_i$	+	$(\log b)\sum_i x_i^2$

The estimated coefficient of determination is given by,

$$\hat{R}^2_{\log y.x} = 1 - \frac{(n-2)s^2_{\log y.x}}{(n-1)s^2_{\log y}}$$

where,

$$s^2_{\log y.x} = \frac{1}{n-2}\{\sum_i (\log y_i)^2 - (\log a \sum_i \log y_i + \log b \sum_i x_i \log y_i)\}$$

The estimated coefficient of correlation is,

$$\hat{R}_{\log y.x} = +\sqrt{\hat{R}^2_{\log y.x}}$$

13.7.3 The $\frac{x}{y}$, x relationship[1/]

For this relationship the estimating equation is of the type,

$$\frac{x_i}{y_c} = a + bx_i$$

The normal equations are,

	(3)	:	(1)	+	(2)
(1)	$\sum_i (\frac{x_i}{y_i})$	=	na	+	$b\sum_i x_i$
(2)	$\sum_i (\frac{x_i^2}{y_i})$	=	$a\sum_i x_i$	+	$b\sum_i x_i^2$

The estimated coefficient of determination is given by,

$$\hat{R}^2_{x/y.x} = 1 - \frac{(n-2)s^2_{x/y.x}}{(n-1)s^2_{x/y}}$$

where,

$$s^2_{x/y.x} = \frac{1}{n-2}\{\sum_i (\frac{x_i}{y_i})^2 - (a\sum_i \frac{x_i}{y_i} + b\sum_i \frac{x_i^2}{y_i})\}$$

1/ The model has a good fit to the empirical data when the values of the variates x_i and x_i/y_i follow arithmetical progressions.

and,

$$s^2_{x/y} = \frac{1}{n-1}\{\sum_i (\frac{x_i}{y_i})^2 - \frac{\left[\sum_i (\frac{x_i}{y_i})\right]^2}{n}\}$$

The estimated coefficient of correlation is,

$$\hat{R}_{x/y.x} = +\sqrt{\hat{R}^2_{x/y.x}}$$

13.7.4 The $\frac{1}{y}$, x relationship

For this relationship the estimating equation is,

$$\frac{1}{y_c} = a+bx_i$$

The normal equations are,

	(3) : (1) + (2)
(1)	$\sum_i (\frac{1}{y_i}) = na + \sum_i x_i$
(2)	$\sum_i (\frac{x_i}{y_i}) = a\sum_i x_i + b\sum_i x_i^2$

etc.

13.7.5 Use of transformations in multiple curvilinear regression

Using logarithms, reciprocals, roots, powers, or some other function of the values of one (or more) of the series may result in reducing a non-linear relation to linear form. Below we give examples of various types of estimating equations,

$$y_c = a+b_1x_{1i}+b_2\log x_{2i}$$

$$y = a+b_1\sqrt{x_{1i}}+b_2x_{2i}$$

$$y = a+b_1\frac{1}{x_{1i}}+b_2x_{2i}$$

$$\log y = a+b_1x_{1i}+b_2x_{2i}$$

etc.

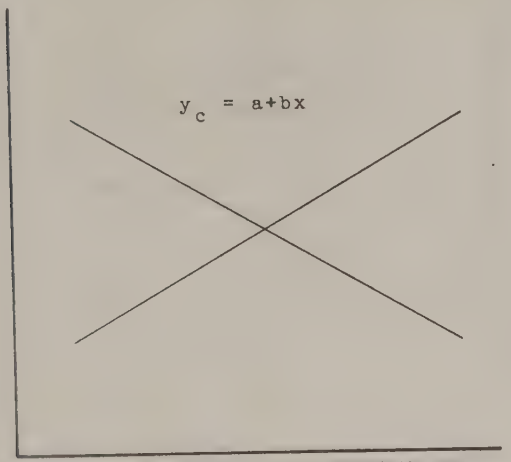


Figure 13.7.5.1a

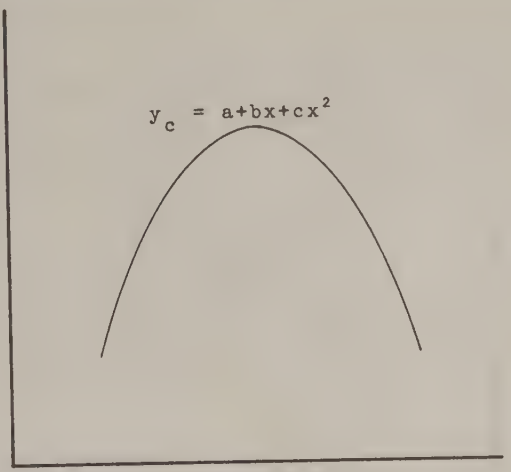


Figure 13.7.5.1b

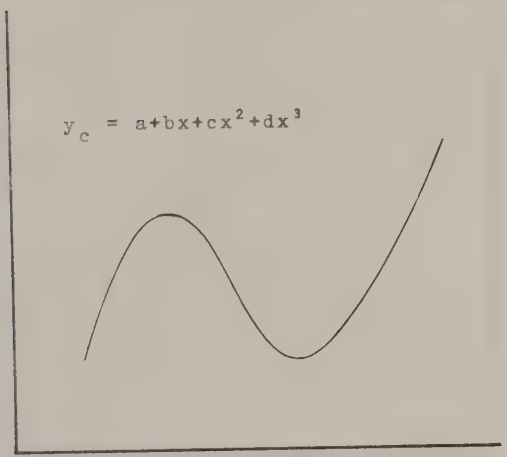


Figure 13.7.5.1c

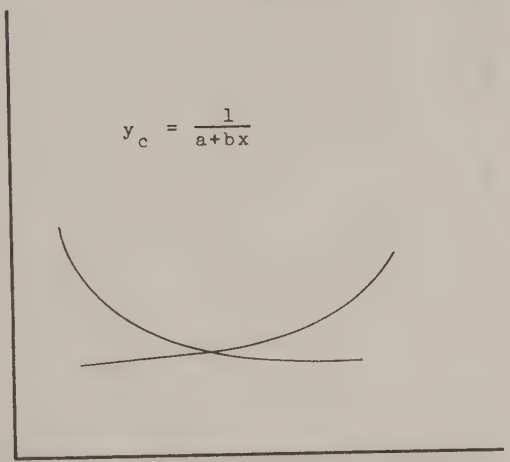


Figure 13.7.5.1d

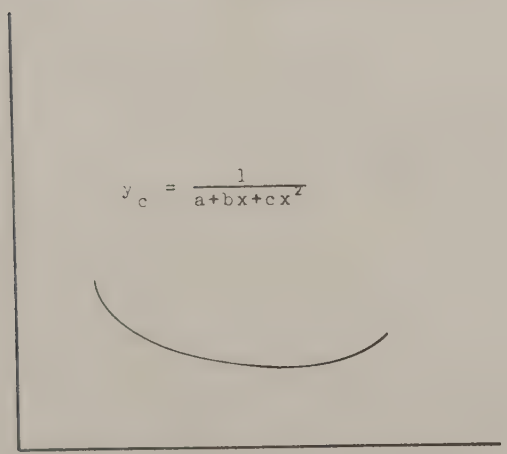


Figure 13.7.5.1e

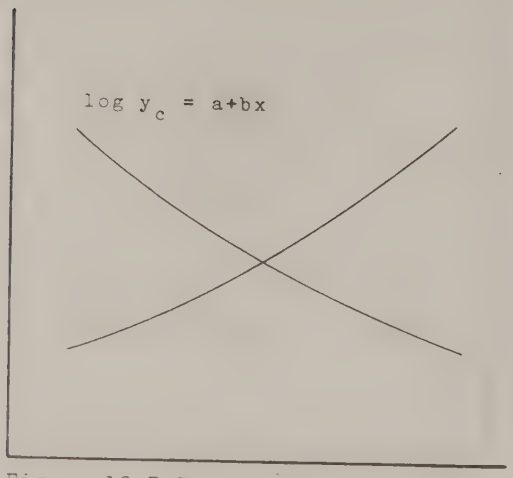


Figure 13.7.5.2a

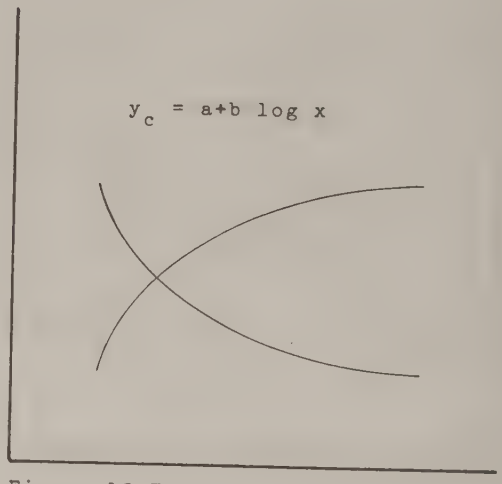


Figure 13.7.5.2b

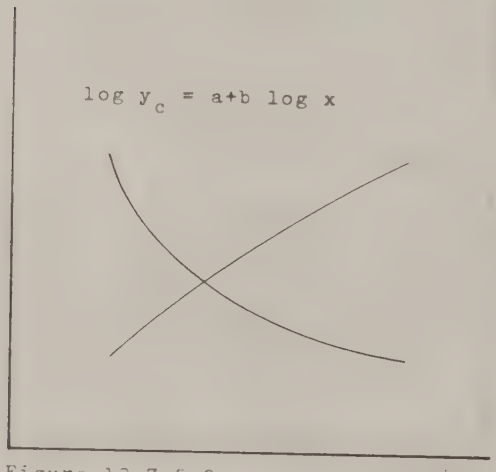


Figure 13.7.5.2c

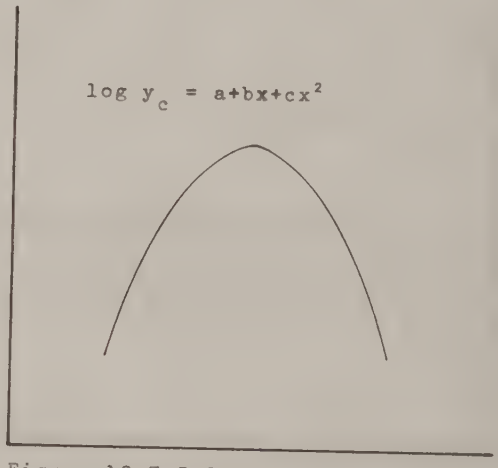


Figure 13.7.5.2d

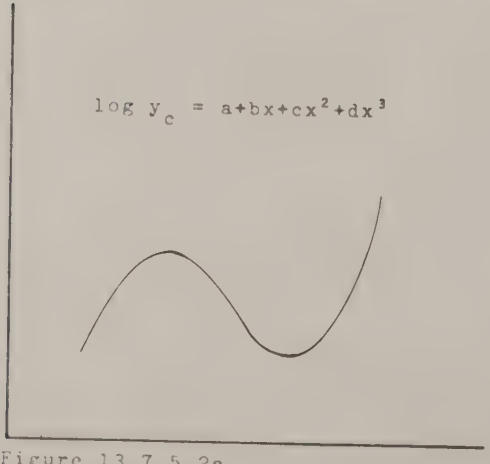


Figure 13.7.5.2e

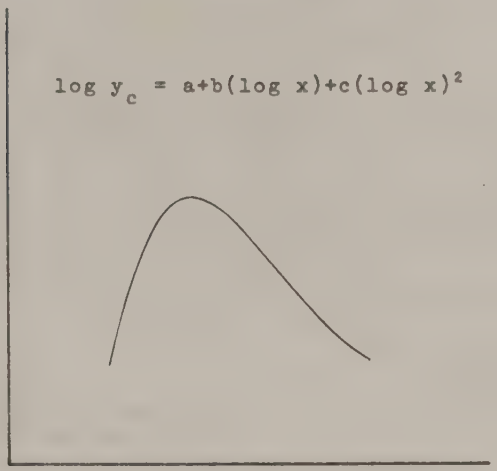


Figure 13.7.5.3a

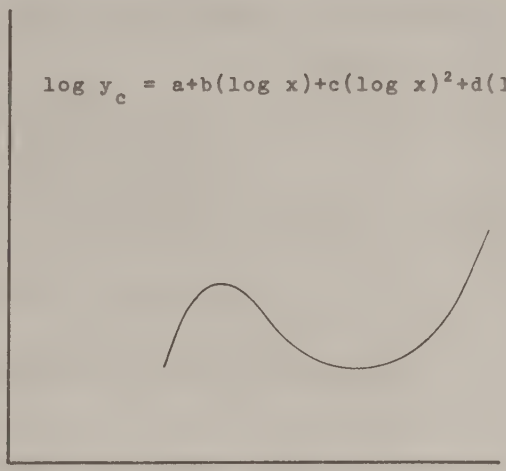


Figure 13.7.5.3b

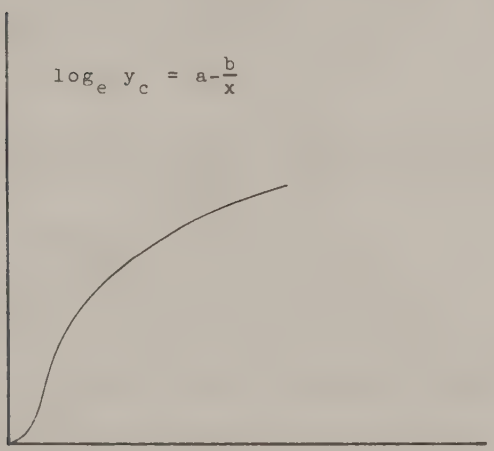


Figure 13.7.5.3c

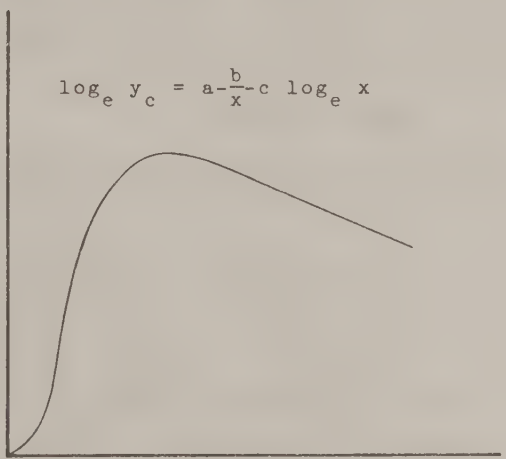


Figure 13.7.5.3d

Curves illustrating a number of different types of mathematical functions

14. METHODOLOGY OF FISHERY TIME SERIES

14.1 Introduction

Regression analysis, described in Chapter 13, gives a static picture of the relationship between the survey variates. The main advantage of this method is that, by applying the proper statistical techniques on the empirical data it permits one to estimate the influence of independent variables on the dependent variable. The factors influencing the dependent variable may also be studied in a dynamic way by comparing the survey variates over a period of successive points in time. The purpose of this chapter is to present some statistical methods describing time series. Specifically, in the field of applied fishery researches, one of the main reasons for attempting to describe a time series by means of a mathematical model is to forecast the future behaviour of the series.

14.2 Introducing time as an explanatory variable

One of the main problems in the field of fishery statistical studies in inland waters in Africa is that of making correct, or at least good, guesses or estimates of what future changes in dynamic magnitudes (fish catch, etc.) will be. In undertaking to solve this problem, it becomes necessary to analyze data over a period of time. Data presenting change over a period of time are known as time series. It should be noted that the fluctuations of a dynamic magnitude are not mathematical totals which vary on a regular schedule but are the net result of many factors which are operating with an unknown weighting system.

14.3 Basic trends[1/]

Basic or secular trends in a dynamic magnitude of a fishing industry may be defined as the underlying tendency to increase or decrease over a period of years. From a statistical point of view all basic trends are not of the same nature:

1. Sometimes the predominating tendency will be a constant amount of growth (which would be shown by a straight line on an arithmetic scale).
2. Sometimes it will be a constant percentage increase (which is shown by a straight line on a semi-logarithmic chart).
3. Sometimes the characteristic tendency is best described by the "logistic" or "S" curve.
4. Sometimes it is a combination of types or a special irregular curve.

It should be noted that if trend movements have been established definitely in the past, and no reason for radical change is apparent, there is statistical justification for projecting the same line (curve) into the future for a year or two.

14.3.1 Basic trend 1 : the straight line

In fitting a trend the first essential is to chart the data, as this is the best way to reveal the movements. Since the most essential thing to know in trend fitting is the general form that the growth curve must assume, statistical models have been developed. By means of such a model a researcher can, for instance, say that at a given time a series shows a certain ratio, or a certain amount, of growth per annum,

---

1/ In the field of economic statistics the time series movements are distinguished by basic or secular trend, periodic or seasonal, cyclical and irregular. For our purposes the basic trends of fishery time series and correlation of the series is considered. For the analysis of economic time series see:

Davis, T.H. (1941), The Analysis of Economic Time Series. Cowles Commission for research in economics. Monograph No.6, Principia Press Inc., Bloomington, Indiana.

Cunnyngham, J. (1963), The Spectram Analysis of Economic Time Series. U.S. Dept. of Commerce, Bureau of the Census, Working Paper No.14, Washington D.C.

and that, if this tendency continues, the trend will reach a certain value at some specified time in the future.

The simplest type of curve is the straight line, which is described by an equation of the type,

$$y_c = a+bt_i$$

in which t (years) is the independent variable and y_c the trend value of the dependent variable.

The method of least squares provides a convenient device to obtain an objective fit of a straight-line trend line to a series of empirical data. As we have discussed (Chapter 13) the method of least squares is a mathematical device which places a line through a series of plotted points in such a way that the sum of the squares of the deviations of the actual points above and below the trend line is at a minimum. For a fitted line the values of a and b can be calculated by solving the normal equations simultaneously,

	(3) : (1) + (2)
(1)	$\sum_i y_i = na + b\sum_i t_i$
(2)	$\sum_i y_i t_i = a\sum_i t_i + b\sum_i t_i^2$

Our manipulations for calculating the values of a and b can be simplified if the origin of t is taken at the first year (to = 0).

Example

The tabulated data (Table 14.3.1.1) give the annual total fish catch (m.tons) at the filling period of a man-made lake. Prepare the respective chart and obtain an objective fitting to the series of empirical data.

Table 14.3.1.1 Annual total fish catch for the first eight years of the filling period of a man-made lake (hypothetical data)

Year	t_i	y_i (m.tons)	Remarks
1965	0	500	$\{\sum_i t_i y_i = 70900\}$
1966	1	800	
1967	2	1200	$\{\sum_i t_i^2 = 140\}$
1968	3	1500	
1969	4	2000	$\{(\sum_i t_i)(\sum_i y_i) = 420000\}$
1970	5	2400	
1971	6	3000	$\{\bar{t} = 3.5, \ \bar{y} = 1875\}$
1972	7	3600	
	28	15000	

$$b = \frac{n\sum_i t_i y_i - (\sum_i t_i)(\sum_i y_i)}{n\sum_i t_i^2 - (\sum_i t_i)^2} = \frac{8\times 70900 - 420000}{8\times 140 - (28)^2} = \frac{147200}{336} = 438.10 \text{ m.tons}$$

It should be noted that b is the slope of the line and expresses the amount of increase in the trend (438.10 m.tons) each year.

$$a = \bar{y} - b\bar{t} = 1875 - 438.10\times 3.5 = 341.65$$

$$y_c = 341.65+438.10t$$

Other things being equal, the expected quantity of fish catch for the years 1973 and 1974 would be,

$$y_{c/73} = 341.65+438.10\times8 = 3846.45 \text{ m.tons}$$

$$y_{c/74} = 341.65+438.10\times9 = 4284.55 \text{ m.tons}$$

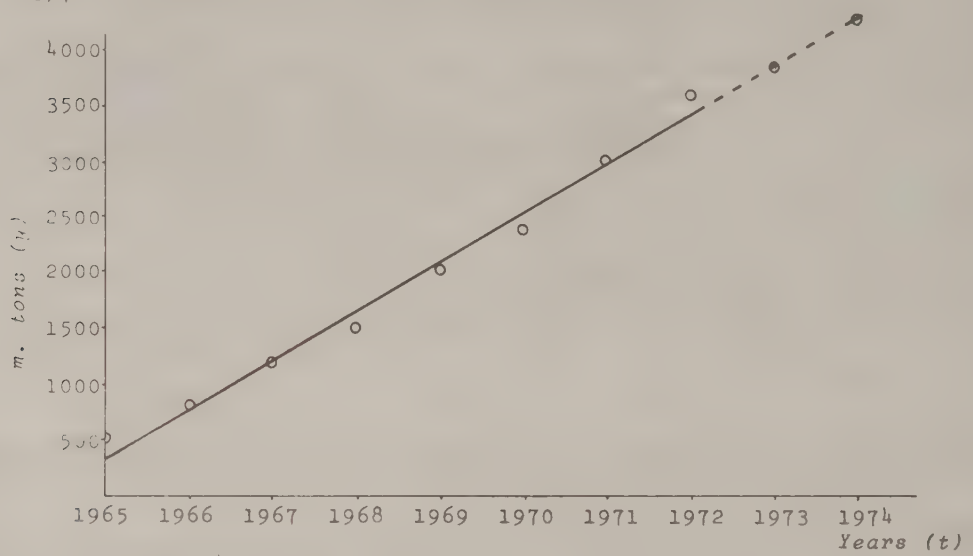


Figure 14.3.1.1

14.3.2 Basic trend 2 : second-degree curve

When a third constant is added to the equation for a straight line, the second-degree curve, which has one bend, is obtained. Because of the bend in the second-degree curve, the slope of the curve is continually changing. Portions of such a curve fitted to a time series may slope upward or downward (or upward in one portion and downward in another) and may be concave upward or concave downward[1/].

The second-degree equation is given by,

$$y_c = a+bt_i+ct_i^2$$

Since we are dealing with three parameters (statistics) in the second-degree curve, a system of three normal equations is required,

	(4) : (1) + (2)+ (3)
(1)	$\sum_i y_i = na +b\sum_i t_i+c\sum_i t_i^2$
(2)	$\sum_i y_i t_i = a\sum_i t_i+b\sum_i t_i^2+c\sum_i t_i^3$
(3)	$\sum_i y_i t_i^2 = a\sum_i t_i^2+b\sum_i t_i^3+c\sum_i t_i^4$

1/ While a straight line indicates a constant amount of increase or decrease, a second-degree curve involves increasing or decreasing amounts of increase or decrease.

If the origin of t can be taken at the middle year, or between the two middle years, (to = 0) the above three normal equations become,

$$\sum_i y_i = na + c\sum_i t_i^2$$

$$\sum_i y_i t_i = b\sum_i t_i^2$$

$$\sum_i y_i t_i^2 = a\sum_i t_i^2 + c\sum_i t_i^4$$

Example

The table below (Table 14.3.2.1) gives the number of active fishing boats at lake L for the period 1965-1973. Prepare the respective chronogram and fit a second-degree curve to the empirical data.

Table 14.3.2.1 Number of active fishing boats at lake L over time, 1965-1973

- hypothetical data -

Year	t_i	No.of fishing boats y_i	Remarks
1965	-4	5200	($\sum_i t_i^2 = 60$)
1966	-3	4600	
1967	-2	6100	($\sum_i t_i^4 = 708$)
1968	-1	6600	
1969	0	6400	
1970	1	6200	($\sum_i y_i t_i = 23300$)
1971	2	8600	
1972	3	8300	($\sum_i y_i t_i^2 = 384500$)
1973	4	7100	
	0	59100	

System of normal equations:

(1) $59100 = 9a+60c$

(2) $23300 = 60b$

(3) $384500 = 60a+708c$

$$b = \frac{23300}{60} = 388.33$$

(1) $59100 = 9a+60c$

(2) $384500 = 60a+354c$

By solving the above system of equations simultaneously we get,

$a = 6772.32, \quad c = -30.85$

$$y'_c = 6772.32+388.33t_i-30.85t_i^2$$

Figure 14.3.2.1

14.3.3 Basic trend 3 : compound-interest curve

Different rates of growth provide curves of varying shapes, but the compound-interest curve trend is readily recognized when plotted on semi-logarithmic paper because on such paper it becomes a straight line,

$$y_c = ab^{t_i}$$

or,

$$\log y_c = \log a + (\log b)t_i$$

The curve of this type indicates a constant ratio of change and therefore a constant ratio of change in the amount of change. Specifically, if b is a positive number greater than one, the trend is upward and the amount of change is undergoing a constant percentage of increase. If b is a positive number smaller than one the trend is downward and the amount of change shows a constant percentage of decrease[1/].

The compound-interest curve is fitted by the method of least squares by substituting logarithms for the actual data. The system of normal equations is given by,

	(3) :	(1) + (2)
(1)	$\sum_i \log y_i$	$= n(\log a) + (\log b)\sum_i t_i$
(2)	$\sum_i (\log y_i)t_i$	$= (\log a)\sum_i t_i + (\log b)\sum_i t_i^2$

Example

The table following (Table 14.3.3.1) gives the number of traps used in a new man-made lake during the period 1967-1971. By using semi-logarithmic paper prepare the respective chronogram and fit a semi-logarithmic curve to the empirical data.

1/ Sometimes data are encountered which, when plotted on semi-logarithmic paper, continue to show curvature, being concave either upward or downward. In such a case we may fit a second-degree curve to the logarithms of the y values,

$$y_c = ab^{t_i}c^{t_i^2} \qquad \text{or,} \qquad \log y_c = \log a + (\log b)t_i + (\log c)t_i^2$$

Table 14.3.3.1 Number of traps used in a new man-made lake during the period 1967-1971

Year	t_i	y_i traps	log y_i	Remarks
1967	-2	62	1.7924	$\{\sum_i t_i^2 = 10\}$
1968	-1	180	2.2553	
1969	0	115	2.0607	
1970	1	190	2.2788	$\{\sum_i t_i(\log y_i) = 1.7547\}$
1971	2	455	2.6580	
			11.0452	

System of normal equations, $(\sum_i t_i = 0)$

$$\sum_i \log y_i = n(\log a)$$

$$\sum_i (\log y_i)t_i = (\log b)\sum_i t_i^2$$

and,

$$\log a = \frac{11.0452}{5} = 2.2090$$

or,

$$a = 161.8080$$

$$\log b = \frac{1.7547}{10} = 0.17547$$

or,

$$b = 1.4979$$

$$\log y_c = 2.2090+0.17547t_i$$

or,

$$y_c = 161.8080\ (1.4979^{t_i})$$

Estimate the expected number of traps to be used in the year 1972:

$$\log y_{c/72} = 2.2090+0.17547\times 3 = 2.73541$$

and,

$$y_{c/72} = 544 \text{ traps}$$

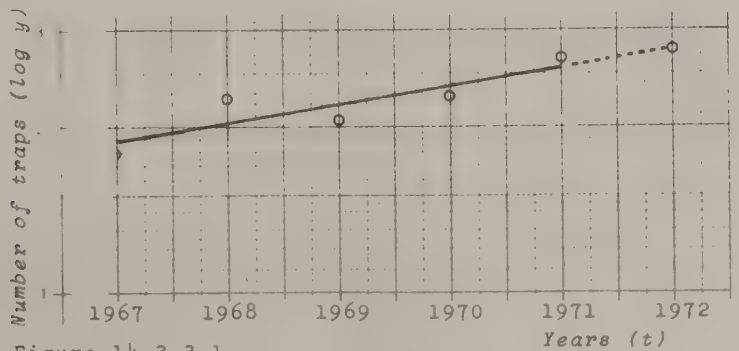


Figure 14.3.3.1

14.3.4 Basic trend 4 : modified exponential curve

Over periods of time, chronological series are not likely to show either a constant amount of change or a constant ratio of change. A survey magnitude over time may have small gains at first, then relatively large gains and then small gains again. Trend curves showing these tendencies towards an 'S' form are known as "S curves".

The modified exponential curve not only describes a trend in which the amount of growth declines by a constant percentage, but also the curve approaches an upper limit, called the asymptote,

$$y_c = k + ab^{t_i}$$

This curve has three constants: k, the asymptote; a, the distance between the value of y_c when t = 0 and the asymptote; and b, the ratio between the successive first differences. Three equations, therefore, are required in order to fit it. This is obtained by dividing the empirical data into three equal sections. The values of y are then totalled for each section. If Σ_1, Σ_2, Σ_3 are the calculated respective subtotals (by section), the values of the constants in the equation are estimated by the formulae[1/],

$$b^n = \frac{\Sigma_3 - \Sigma_2}{\Sigma_2 - \Sigma_1}$$

$$a = \frac{b-1}{(b^n-1)^2}(\Sigma_2 - \Sigma_1)$$

$$k = \frac{1}{n}\left(\frac{\Sigma_1\Sigma_3 - \Sigma_2^2}{\Sigma_1 + \Sigma_3 - 2\Sigma_2}\right)$$

where n is the number of years in each section.

Example

The table below (Table 14.3.4.1) gives catches per unit of effort of the canoe fishery at lake L for the period 1968-1973. By using the empirical data prepare the respective chronogram and fit a modified exponential curve.

Table 14.3.4.1 Estimated catches per unit of effort (lb/canoe/day) of the canoe fishery at lake L over time, 1968-1973

Year	t	y (lb/canoe /day)	Σ_i	y increment	y(%) of preceeding increment
1968	1	40		...	...
1969	2	56	$\Sigma_1 = 96$	16	...
1970	3	68		12	75
1971	4	77	$\Sigma_2 = 145$	9	75
1972	5	84		7	78
1973	6	89	$\Sigma_3 = 173$	5	77

1/ The equation is not obtained by the method of least squares, but it is so fitted that the three partial totals of the trend values are the same as those of the original data.

$$b^2 = \frac{173-145}{145-96} = 0.5714$$

and,

$$b = \sqrt{0.5714} = 0.7559$$

$$a = \frac{(0.7559-1)}{(0.5714-1)^2} \times (145-96) = -65.1111$$

$$k = \frac{1}{3}\left(\frac{96\times173-145^2}{96+173-2\times145}\right) = 105.1667 \text{ lb/canoe/day (upper asymptote value)}$$

$$y_c = 105.1667-(65.1111)(0.7559)^{ti}$$

Estimate the expected value of y for the year 1974.

$$= 105.1667-(65.1111)(0.7559)^7$$

$$= 105.1667-65.1111\times0.1410 = 95.9860 \text{ lb/canoe/day}$$

Figure 14.3.4.1

14.3.5 Basic trend 5 : Gompertz curve[1/]

There is of course no limit to the number of kinds of trends which may exist. The Gompertz curve describes a series in which the natural values of the trend would show a declining ratio of increase, but the ratio does not decrease by either a constant amount or a constant percentage. Thus, it describes a trend in which the growth increments of the logarithms are declining by a constant percentage. The equation for the Gompertz curve is,

$$y_c = ka^{b^{ti}}$$

which may be put into logarithmic form as,

$$\log y_c = \log k+(\log a)b^{ti}$$

1/ See also, G.P. Bazigos, Yield Indices in Inland Fisheries with Special Reference to Volta Lake, St.S./3, FIO/SF/GHA/10, September 1971

This curve (and also the modified exponential curve in section 14.3.4) has an upper and lower asymptote.

By using the notation of section 14.3.4, the formulae for estimating the constants of the equation are,

$$b^n = \frac{\Sigma_3(\log y)-\Sigma_2(\log y)}{\Sigma_2(\log y)-\Sigma_1(\log y)}$$

$$\log a = \frac{b-1}{(b^n-1)^2}\times\{\Sigma_2(\log y)-\Sigma_1(\log y)\}$$

$$\log k = \frac{1}{n}\left(\frac{\Sigma_1(\log y)\Sigma_3(\log y)-\{\Sigma_2(\log y)\}^2}{\Sigma_1(\log y)+\Sigma_3(\log y)-2\Sigma_2(\log y)}\right)$$

Example

In the table below (Table 14.3.5.1) quarterly sample catch data at a man-made lake L were inferred to the yearly totals. Prepare the respective chronogram and fit a Gompertz curve to the empirical data.

Table 14.3.5.1 Estimated fish catch (m. tons) at man-made lake L over time

Year	Quarter	y_i Estimated yearly totals	$\log y_i$	Remarks
1965	1 2 3 4	2772 2554 2264 1618	3.4428 3.4072 3.3549 3.2090	$\Sigma_1(\log y) = 13.4139$
$\Sigma_1(\log y)$	...	...	13.4139	
1966	1 2 3 4	3181 3901 4642 4116	3.5026 3.5912 3.6667 3.6145	$\Sigma_2(\log y) = 14.3750$
$\Sigma_2(\log y)$	...	...	14.3750	
1967	1 2 3 4	5688 5927 6051 5968	3.7550 3.7728 3.7818 3.7758	$\Sigma_3(\log y) = 15.0854$
$\Sigma_3(\log y)$	...	...	15.0854	

$$b^4 = \frac{15.0854-14.3750}{14.3750-13.4139} = 0.7392$$

and,

$$\log b = \frac{1}{4}\log 0.7392 = -0.03281$$

and,

$$b = 0.9272$$

$$\log a = \frac{-0.0728}{0.0680}\times(14.3750-13.4139) = -1.0289$$

and,

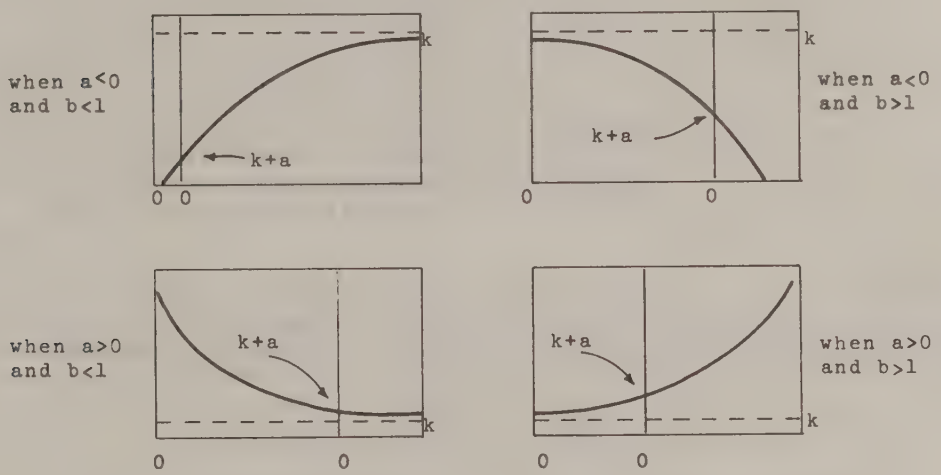
$a = 0.09356$

$$\log k = \frac{1}{4}\left(\frac{13.4139 \times 15.0854 - 14.3750^2}{13.4139 + 15.0854 - 2 \times 14.3750}\right) = 4.2746$$

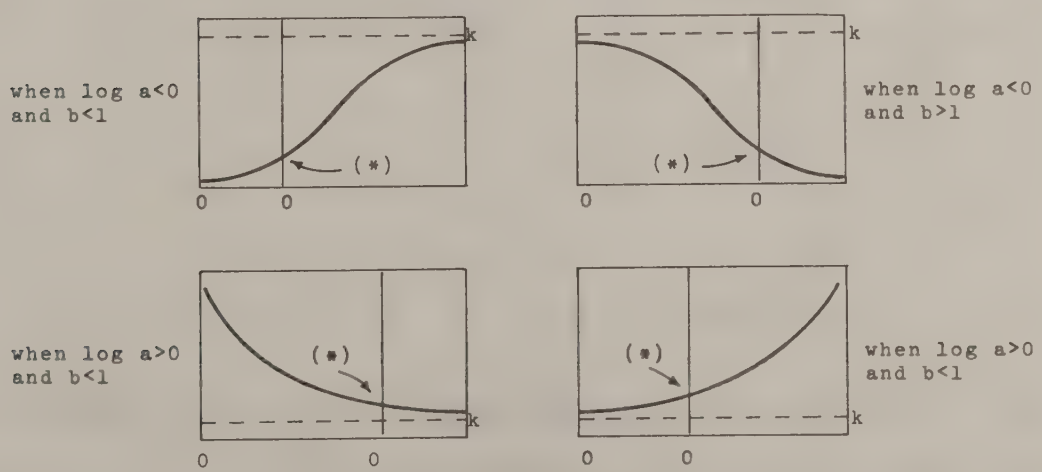
and,

$k = 18819$ m.tons (upper asymptote value)

$\log y_c = 4.2746 - 1.0289(0.9272^{t_i})$



Four forms of the Modified Exponential Curve $y_c = k + ab^{t_i}$



Four forms of the Gompertz Curve, $y_c = ka^{b^{t_i}}$. The vertical values marked (*) are antilog ($\log^c k + \log a$)

14.4 Correlation of fishery time series

In the field of applied fishery researches the problem of correlation of time series arises in the following two cases:

1. Data of specified reference periods (annual, quarterly, etc.) for one series are regularly known or made available (variate x) before the corresponding data for another, closely correlated series (variate y). In such a situation, if the correlation is high a useful estimate may be made for the series which is not so promptly available.

2. A time series (variate x) may move ahead of a second series (variate y) and which may therefore be used to forecast the second series. When such a situation occurs, the earlier series is said to "lead" the other series. Also, the later-moving series is said to "lag" the earlier-moving series.

   At Lake Volta (Ghana) a lead-lag relationship was obtained between total alkalinity and estimated total annual fish catch. The leading series was the total alkalinity and the lagging series the total annual fish catch[1/]. At Lake Kainji (Nigeria) a lead-lag relationship was obtained between the catches per unit of effort (experimental gillnet survey) and estimated total yearly fish catch. The leading series was the catch per unit of effort and the lagging series was the total yearly fish catch[2/].

Example

The table below (Table 14.4.1) gives the estimated total yearly fish catch (x) and total commercial fish catch[3/] (y) at lake L over time. Estimate the regression equation of y on x and the expected value of y for the year 1971.

Table 14.4.1 Estimated total yearly fish catch and commercial fish catch at lake L over time

Year	t_i	x_i (fish catch)	y_i (commercial fish catch)	Remarks
1965	1	1020	900	($\sum_{i=1}^{6} x_i = 5910$, $\sum_{i=1}^{6} y_i = 5330$)
1966	2	1010	930	
1967	3	1050	940	($\sum_{i=1}^{6} x_i y_i = 5260400$)
1968	4	960	870	
1969	5	920	820	($\sum_{i=1}^{6} x_i^2 = 5833500$, $\sum_{i=1}^{6} y_i^2 = 4744700$)
1970	6	990	870	
1971	7	1080	...	($\bar{x} = 985.00$, $\bar{y} = 888.33$)
		6990	(5330)	

1/ G.P. Bazigos, Yield Indices in Inland Fisheries with Special Reference to Volta Lake, St.S./3, FIO/SF/GHA/10, September 1971.

2/ G.P. Bazigos, Recent Trends in the Yield Pattern of Kainji Lake, St.S./3, UNDP/SF/NIR/24, April 1974.

3/ Commercial fish catch is the quantity of fish catch which enters the commercial channel.

Estimated linear correlation coefficient, r_{xy},

$$r_{xy} = \frac{6\times5260400-5910\times5330}{\sqrt{(6\times5833500-5910^2)(6\times4744700-5330^2)}} = 0.9445$$

Estimated linear regression of y on x,

$$y = a+bx$$

$$b = \frac{6\times5260400-5910\times5330}{(6\times5833500-5910^2)} = 0.8519$$

The estimated b = 0.8519 indicates that on an average basis, out of 1000 m.tons yearly total fish catch, 852 m.tons enter the commercial channel.

$$a = 888.33-0.8519\times985.00 = 49.209$$

and,

$$y_c = 49.209+0.8519x_i$$

Estimated commercial fish catch for the year 1971,

$$y_{c/71} = 49.209+0.8519\times1080 = 969.261 \text{ m.tons}$$

Example

In man-made lake L a lead-lag relationship exists in the movements of the following two parallel magnitudes: catches per unit of effort (experimental gillnet survey, leading variate x) and total yearly fish catch (catch assessment survey, lagging variate y). The average lead-lag time difference between the two time series is of one year. By using the tabulated data (Table 14.4.2) estimate the regression equation of y on x and the expected quantity of fish catch for the year 1974.

Table 14.4.2 Estimated values of two parallel magnitudes over time

Variate x (kg/CUE)			Variate y (m. tons/ fish catch)			Remarks
Year	t_i	lbs	Year	t_i	m. tons	
1968	1	80	1969	2	5000	($\sum_{i=1}^{5} x_i = 365$, $\sum_{i=2}^{6} y_i = 21900$)
1969	2	100	1970	3	6100	
1970	3	70	1971	4	4200	($\sum_{i=1,2}^{5,6} x_i y_i = 1684500$)
1971	4	60	1972	5	3500	($\sum_{i=1}^{5} x_i^2 = 27925$, $\sum_{i=2}^{6} y_i^2 = 101710000$)
1972	5	55	1973	6	3100	
1973	6	48	1974	7	...	($\bar{x} = 73.00$, $\bar{y} = 4380$)
		413			(21900)	

Estimated linear correlation coefficient, r_{xy},

$$r_{xy} = \frac{5\times1684500-365\times21900}{\sqrt{(5\times27925-365^2)(5\times101710000-21900^2)}} = 0.9968$$

Estimated linear regression of y on x,

$$y = a+bx$$

$$b = \frac{5\times1684500-365\times21900}{(5\times27925-365^2)} = 67.03$$

$$a = 4380-67.03\times73 = -513.19$$

$$y_c = -513.19+67.03x$$

Expected fish catch for the year 1974,

$$y_{c/74} = -513.19+67.03\times48 = 2704.250 \text{ m.tons}$$

14.5 Multiple estimating equations

Sometimes it is desirable to express the relationship existing between more than two variables over time by means of a multiple estimating equation. For example, at Lake Nasser[1/] it was found that, on a yearly basis, changes in the total commercial landings (variate y) were highly correlated (positively) with changes in the total number of fishing boats (variate x_1) and changes in the total surface area of the lake (variate x_2). A multiple regression model (linear) of commercial landings on fishing boats and surface area of the lake was established,

$$y_c = a+b_1x_{1i}+b_2x_{2i}$$

Specifically, in order to assess the linearity of the model, the correlation matrix was produced. The table below (Table 14.5.1) gives the estimated linear correlation coefficients between the pairs of variables under study.

Table 14.5.1 Correlation matrix

	y	x_1	x_2
y	$r_{yy} = 1.000000$	$r_{yx_1} = 0.989974$	$r_{yx_2} = 0.999018$
x_1		$r_{x_1x_1} = 1.000000$	$r_{x_1x_2} = 0.982728$
x_2			$r_{x_2x_2} = 1.000000$

By using the method of least squares (see Chapter 13) an objective fit of the mathematical model trend to the series of the empirical data was obtained. The estimated multiple equation was,

$$y_c = -1045.68+1.3843x_{1i}+1.8449x_{2i}$$

15. FISH YIELD INDICES

15.1 Introduction

Fish yield indices can be considered as a simple statistical method to portray changes in space and time of the total catches obtainable from inland water fisheries.

1/ G.P. Bazigos, The Yield Pattern at Lake Nasser (Arab Republic of Egypt), St.S./1, UNDP/SF/EGY/66/558, September 1972.

The estimated index[1/] can be used, among other things, as an indicator for measuring changes in the abundance of fish over time (see section 15.2.4).

15.2 The construction of the index

Among the problems encountered in the construction of the index are,

1. The population of species.
2. Selection of the base.
3. The method of combining the data.

15.2.1 The population of species

In the first place, an index number of fish yield in inland waters must, of necessity, refer to a given set of commercially important species (groups of species). However, it is necessary to recognize that on a long-term basis the set of commercially important species is not a constant list (specifically in man-made lakes) and this inconsistency is almost universally met with in other kinds of fisheries. The greater the field of species involved the greater the discrepancy is likely to be.

It will be worthwhile to introduce special terminology to emphasize the above mentioned situation,

1. Totality of species: refers to all commercially important species (groups of species) of the several time periods 0, 1, 2, etc.
2. Binary species: refers to the species (groups of species) found in both of the two specified periods 0 and 1.
3. Unique species: refers to the species (groups of species) found in either of the two lists of a binary comparison but not in both.

It is clear from the above discussion that neither "binary species" nor "unique species" may be specified for a given period until a second period has been named for comparison purposes.

As an example, at Volta Lake (Ghana) fish catch data by species have been collected by the Catch Assessment Survey (CAS) for 16 commercially important species or groups of species, the population of species being kept fixed over time. However, analysis of the results of the survey showed that seven species out of sixteen comprise about 80 percent of the total catch and these species were the most economically important. The species are given in the table below (Table 15.2.1.1). It was decided that the species coverage of the index be confined to the seven most important species, plus other species.

Table 15.2.1.1 Species coverage of the yield index (Volta Lake)

Species (groups of species)	
C.No.	Name
1	Lates
2	Tilapia
3	Labeo
4	Hydrocyon
5	Alestes
6	Citharinus
7	Distichodus
8	Others

1/ This method has been developed by the author and successfully applied in the field of applied fishery researches in inland waters of Africa: G.P. Bazigos, Yield Indices in Inland Fisheries with Special Reference to Volta Lake, St.S./3, FIO/SF/GHA/10, September 1971. G.P. Bazigos, Yield Indices at Lake Tanganyika (Burundi) Working Paper No.2, UNDP/SF/BDI/70/508, September 1973.

15.2.2 Selection of the base

Regardless of the formula employed to combine the data (see sections following), it is customary to select some period of time as 100 percent with which to compare the other index numbers. A good base period is a normal period as far as movements of the survey magnitude are concerned. The choice of the proper base period is not always an easy task. For example, in man-made lakes the filling of the reservoir must be considered as a transition period and it can in no way contribute to the base period. Further, a high level of productivity exists in newly impounded reservoirs for perhaps the first year or two. This is due to phytobiomass resulting from the flooded vegetation which can be viewed as a "flash in the pan". It exists for a relatively short time and will probably have little bearing on the eventual level at which the reservoir productivity will stabilize. In such cases the proper base period can be obtained by a careful study of the fluctuations of the survey magnitude over time. At Volta Lake all the properties of a normal year period were inherent in 1970[1/].

15.2.3 The data for the indices

The raw material for the construction of the yield indices is the sample estimates of the CAS[2/]. By this method, the precision of the estimated indices can be assessed. In addition, it becomes possible to infer whether observed variations of the estimated indices are due to sampling fluctuations, or whether they actually reflect real changes of the magnitudes under estimation.

15.2.4 Yield indices defined

In order to satisfy the demands of the various consumers of catch statistics two kinds of indices have been introduced,

1. Yield indices, type 1: These indices are constructed to measure at regular intervals (yearly) changes in the total volume of catch by fishery and by region and for the lake as a whole. It is believed that the indices of type 1 satisfy the general demands of the respective governments and of other organizations by showing the trends of the fish catch of the various domains of a fishing industry.
2. Yield indices, type 2: The series of these indices has been introduced in order to serve the specific demands of biologists and other people responsible for fisheries management at the lakes. Specifically, two groups of indices are constructed,

   Group 2a: The indices show changes in the abundance of fish over time. These indices are obtained by dividing the values of the yield indices of type 1 by the corresponding values of the fishing effort index[3/].

   Group 2b: Specific indices are constructed showing changes in space and time within species (weight, number) and in average weight per fish.

1/ The filling of the reservoir took place during the period 1964-1968 and the "surprise year" due to peculiarities of the new reservoir was the year 1969.

2/ G.P. Bazigos, The Design of Fisheries Statistical Surveys - Inland Waters, FAO Fish.Tech.Pap. FIPS/T135, May 1974.

3/ The fishing effort index measures changes of the level of fishing operations in space and time and by fishery. The raw material for the construction of the index is the results of the CAS. The base period of the index is matched with the base period of the yield indices of type 1.

The estimated matrix of indices of types 1 and 2 provide information not only of changes in the individual magnitudes but also of "structural" variations in the population under investigation.

15.2.5 The statistical formulae of estimation

To construct yield indices the method of simple aggregates is used. It is apparent that the influence which a fishery or species etc. exerts on the estimated simple aggregate index depends on the estimated total fish catch (CAS) of the respective statistical magnitudes. For example, if there are two fisheries at a fishing industry ($k = 1, 2$) and the coverage of the index, as far as species is concerned, is s ($i = 1, 2, \ldots s$), then the estimated total fish catch for the years 1 and 0 (base period) is given respectively by,

$$ {}_1\hat{Y} = \sum_{ki} {}_1\hat{Y}_{ki} \qquad \text{and} \qquad {}_0\hat{Y} = \sum_{ki} {}_0\hat{Y}_{ki} $$

where,

${}_1\hat{Y}_{ki}$ and ${}_0\hat{Y}_{ki}$: estimated total fish catch within species i in the k^{th} fishery in the years 1 and 0 respectively (results of CAS).

The estimated yield index (type 1) is,

$$ \hat{I}^{(1)}_{1/0} = \sum_{ki} {}_1\hat{Y}_{ki} \div \sum_{ki} {}_0\hat{Y}_{ki} $$

or,

$$ \hat{I}^{(1)}_{1/0} = \frac{{}_1\hat{Y}}{{}_0\hat{Y}} \times 100 = \ldots (\%) $$

The magnitudes which are involved in the calculation of the index are estimates based on the results of the CAS, i.e. $\hat{I}^{(1)}_{1/0}$ is the ratio of two estimates. In such a case, the sampling error of the estimated index can be calculated by using the proper estimator. If $v({}_1\hat{Y})$ and $v({}_0\hat{Y})$ are the estimated variances of ${}_1\hat{Y}$ and ${}_0\hat{Y}$ and $cv({}_1\hat{Y})$, $cv({}_0\hat{Y})$ their respective relative sampling errors,

$$ \{cv({}_1\hat{Y}) = \sqrt{v({}_1\hat{Y})} \div {}_1\hat{Y}, \quad cv({}_0\hat{Y}) = \sqrt{v({}_0\hat{Y})} \div {}_0\hat{Y} $$

an approximate formula for the estimated relative error of $\hat{I}^{(1)}_{1/0}$ is given by,

$$ cv\{\hat{I}^{(1)}_{1/0}\} = \sqrt{cv^2({}_1\hat{Y}) + cv^2({}_0\hat{Y}) - 2\hat{\rho}\, cv({}_1\hat{Y})\, cv({}_0\hat{Y})} $$

where $\hat{\rho}$ is the estimated linear correlation coefficient between ${}_1\hat{Y}$ and ${}_0\hat{Y}$.

The technique discussed above can be used for estimating the various indices of type 2.

The following tables (Tables 15.2.5.1, 15.2.5.2) show the estimated yield indices of type 1 (unadjusted fish yield indices) and type 2 (group 2a) (adjusted fish yield indices for changes in fishing effort) at Lake Tanganyika (Burundi). Also, the prepared chart units (15.2.5a, b, c) portray the trends over time of the estimated adjusted indices. Fish yield indices have been prepared by major group of species, by fishery and for the industry as a whole (survey period 1966-1971, 1966 = 100, base year).

Estimated fish yield indices - type 1

Table 15.2.5.1 Estimated unadjusted fish yield indices (%) by major group of species, by fishery and for the industry as a whole, 1966-1971, Lake Tanganyika (Burundi)

- (%) -

Fishery/major group of species	1966	1967	1968	1969	1970	1971
0. Total industry						
a. Total fish catch	100.00	73.58	66.04	89.80	77.43	92.12
b. Clupeids	100.00	70.92	60.28	82.92	64.99	93.29
c. Predators	100.00	123.43	174.08	218.82	192.31	70.18
1. Traditional fishery						
a. Total fish catch	100.00	65.66	56.92	62.34	40.33	59.06
b. Clupeids	100.00	65.89	55.65	62.41	40.71	59.73
c. Predators	100.00	46.61	163.56	56.80	8.47	2.54
2. Artisanal fishery						
a. Total fish catch	100.00	73.28	56.19	192.81	141.51	183.52
b. Clupeids	100.00	74.81	55.79	191.80	136.25	184.11
c. Predators	100.00	37.04	65.74	216.67	265.74	169.44
3. Industrial fishery						
a. Total fish catch	100.00	94.08	96.09	90.73	103.91	115.26
b. Clupeids	100.00	83.07	77.65	61.02	83.28	124.49
c. Predators	100.00	153.15	194.99	250.08	214.54	65.75

Estimated fish yield indices - type 2 (group 2a)

Table 15.2.5.2 Estimated adjusted fish yield indices (%) by major group of species and by fishery, 1966-1971, Lake Tanganyika (Burundi)

- (%) -

Fishery/major group of species	1966	1967	1968	1969	1970	1971
1. Traditional fishery						
a. Total fish catch	100.00	65.66	56.92	73.37	44.90	65.75
b. Clupeids	100.00	65.89	55.65	73.50	45.32	66.49
c. Predators	100.00	46.61	163.56	66.85	9.43	2.83
2. Artisanal fishery						
a. Total fish catch	100.00	65.66	50.34	75.45	54.30	70.42
b. Clupeids	100.00	67.03	49.99	75.03	52.28	70.65
c. Predators	100.00	33.19	58.90	84.79	101.97	65.02
3. Industrial fishery						
a. Total fish catch	100.00	117.60	94.21	82.48	74.22	76.84
b. Clupeids	100.00	103.84	76.13	55.47	59.49	82.99
c. Predators	100.00	191.44	191.17	227.35	153.24	43.83

LAKE TANGANYIKA (BURUNDI)

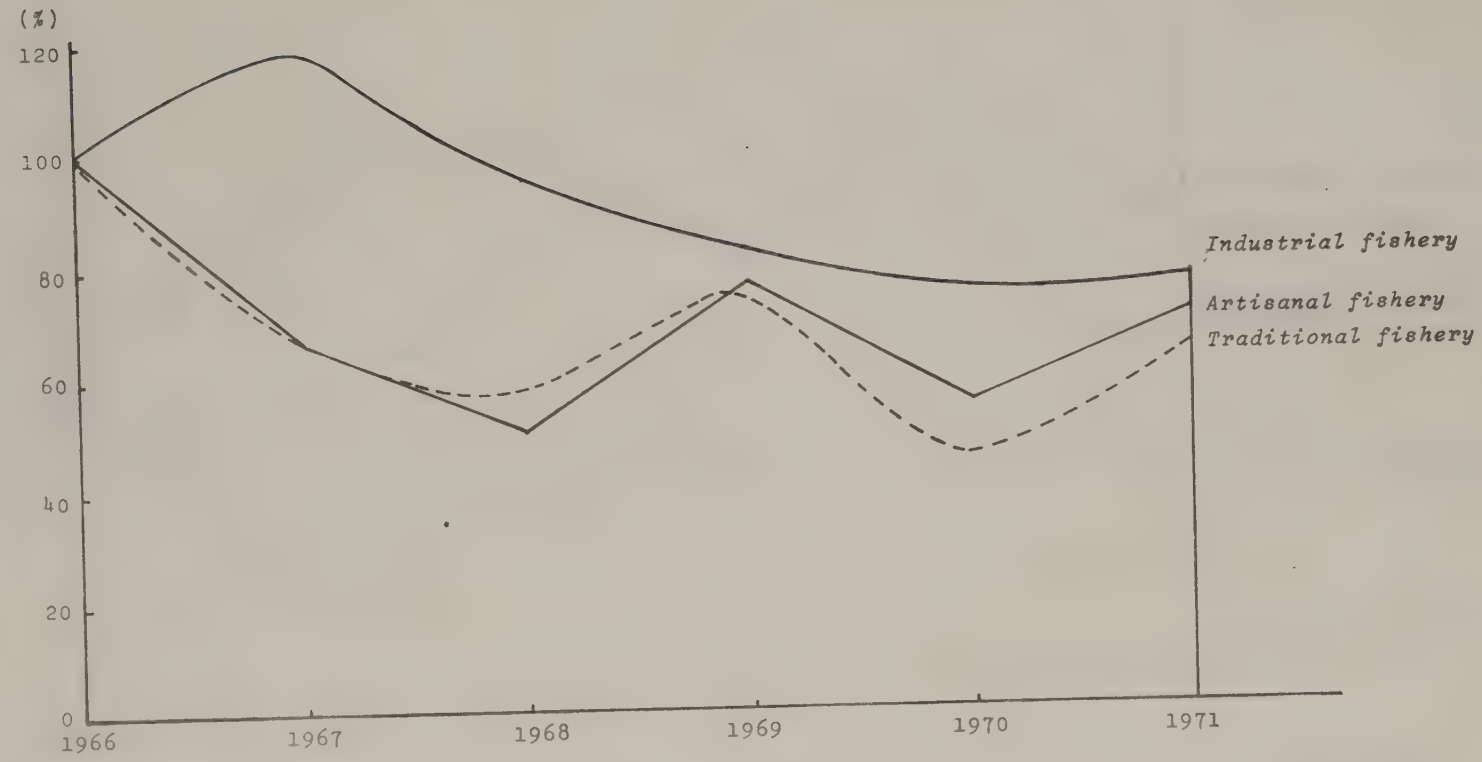


Figure 15.2.5a Estimated adjusted fish yield indices of total fish catch by fishery (1966 = 100)

LAKE TANGANYIKA (BURUNDI)

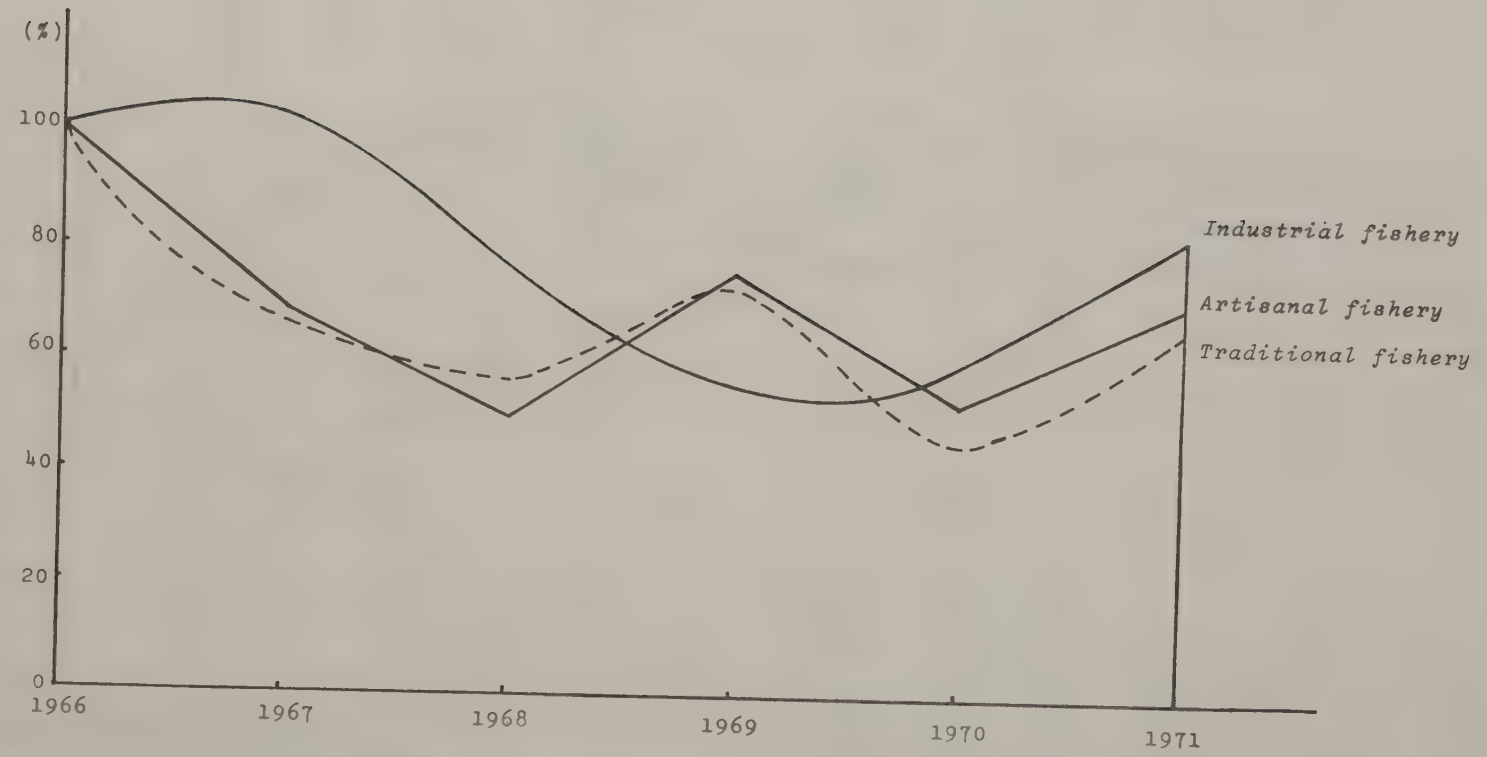


Figure 15.2.5b Estimated adjusted fish yield indices of Clupeid catches by fishery (1966 = 100)

LAKE TANGANYIKA (BURUNDI)

Figure 15.2.5c Estimated adjusted fish yield indices of predator catches by fishery (1966 = 100)

APPENDIX I - MATHEMATICS OF SIMULTANEOUS EQUATIONS

1. SIMULTANEOUS LINEAR EQUATIONS IN TWO UNKNOWNS

The general expression for a system of a pair of simultaneous equations is,

$$\begin{array}{|ll|l} (1) & a_1x+b_1y = c_1 & , \quad a_1 \neq 0 \text{ or } b_1 \neq 0 \\ (2) & a_2x+b_2y = c_2 & , \quad a_2 \neq 0 \text{ or } b_2 \neq 0 \end{array}$$

The method of solution is to transfer the above system into an equivalent system whose solution is obvious.

Example

Solve the system,

$$2x-5y = 19$$
$$3x+4y = -6$$

To eliminate x, we multiply the first equation by 3 and the second equation by -2. This gives us,

$$6x-15y = 57$$

$$-6x-8y = 12$$

Adding we get,

$$-23y = 69 \quad \text{and} \quad y = -\frac{69}{23} = -3$$

Putting y = -3 in the first equation of the system and solving for x we have,

$$2x-5(-3) = 19$$

or, $2x+15 = 19$

or, $2x = 19-15$

or, $2x = 4$, and $x = 2$

Hence the solution is the pair (2, -3)

2. SIMULTANEOUS LINEAR EQUATIONS IN THREE UNKNOWNS

The general expression for a system of three simultaneous equations is,

$$\begin{array}{|lll|l} & (1) & a_1x+b_1y+c_1z = d_1 & \text{Not all } a_1, b_1, c_1 = 0 \\ (A) & (2) & a_2x+b_2y+c_2z = d_2 & \text{Not all } a_2, b_2, c_2 = 0 \\ & (3) & a_3x+b_3y+c_3z = d_3 & \text{Not all } a_3, b_3, c_3 = 0 \end{array}$$

A solution of (A) is an ordered triple (xyz) which satisfies all three equations.

Example

Solve the system,

(1)	2x-y+3z = -9
(2)	x+3y-z = 10
(3)	3x+y-z = 8

First we eliminate x between the (1) and (2) equations and between (1) and (3), i.e.,

(1)	2x-y+3z = -9	(-1)	-2x+y-3z = 9
(2)	x+3y-z = 10	(2)	2x+6y-2z = 20
			7y-5z = 29

(1)	2x-y+3z = -9	(-3)	-6x+3y-9z = 27
(3)	3x+y-z = 8	(2)	6x+2y-2z = 16
			5y-11z = 43

The obtained results with the unchanged first equation, is the equivalent system,

(1)	2x-y+3z = -9
(2)	7y-5z = 29
(3)	5y-11z = 43

Next we eliminate y from the (2) and (3) equations, e.g.

(2)	7y-5z = 29	(-5)	-35y+25z = -145
(3)	5y-11z = 43	(7)	35y-77z = 301
			-52z = 156, and z = -3

Putting z = -3 in the equation (2) we get,

$7y-5(-3) = 29$, and $y = 2$

Putting y = 2 and z = -3 in the first equation we get,

$2x-2-9 = -9$, and $x = 1$

Hence the solution is (1, 2, -3)

3. DETERMINANTS

Let A be the 2×2 matrix $\begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}$. Then we define the expression $a_1b_2-a_2b_1$ to be the determinant of A and write,

$$\det A = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = a_1b_2-a_2b_1$$

Let A be the 3×3 matrix $\begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix}$

If we strike out in A a row and column containing any element, we are left with a 2×2 matrix whose determinant has already been defined. This determinant is called the minor of the corresponding element.

We attach an algebraic sign to each minor,

$\{(-1)^{k+\lambda}$, where suffixes k, λ indicate the line and column in which the element is located in A$\}$

The product of the minor determinant, multiplied by the sign, is called the cofactor of the corresponding element.

The cofactor of any element is denoted by the corresponding capital letter, for instance the cofactor of a_1 is A_1. To define the determinant, we now consider the first row and define,

$$\det A = a_1A_1+b_1B_1+c_1C_1$$

$$= a_1(b_2c_3-b_3c_2)-b_1(a_2c_3-a_3c_2)+c_1(a_2b_3-a_3b_2)$$

Example

Find,

$$\det \begin{pmatrix} 1 & -2 & 3 \\ 4 & 1 & -1 \\ 1 & 2 & 1 \end{pmatrix} = \begin{vmatrix} 1 & -2 & 3 \\ 4 & 1 & -1 \\ 1 & 2 & 1 \end{vmatrix}$$

Choosing the first row, we find that the cofactors are,

$$(-1)^{1+1}\begin{vmatrix} 1 & -1 \\ 2 & 1 \end{vmatrix} = +\{1-(-2)\} = 3$$

$$(-1)^{1+2}\begin{vmatrix} 4 & -1 \\ 1 & 1 \end{vmatrix} = -\{4-(-1)\} = -5$$

$$(-1)^{1+3}\begin{vmatrix} 4 & 1 \\ 1 & 2 \end{vmatrix} = +(8-1) = 7$$

So the answer is,

$$(1)\times(3)+(-2)\times(-5)+(3)\times(7) = 34$$

The same result is obtained if we choose the first column. The cofactors are,

$$(-1)^{1+1}\begin{vmatrix} 1 & -1 \\ 2 & 1 \end{vmatrix} = 3$$

$$(-1)^{2+1}\begin{vmatrix} -2 & 3 \\ 2 & 1 \end{vmatrix} = -\{(-2)-(6)\} = 8$$

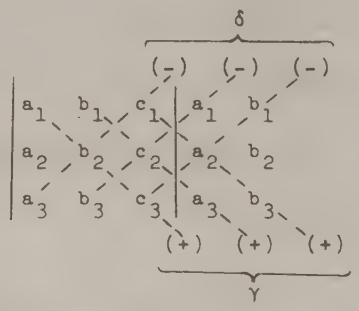
$$(-1)^{3+1}\begin{vmatrix} -2 & 3 \\ 1 & -1 \end{vmatrix} = +\{2-(3)\} = -1$$

So the answer is,

$$(1)\times(3)+(4)\times(8)+(1)\times(-1) = 34$$

Determinants of square matrices of higher dimension are defined inductively in a similar fashion. For the 4×4 matrix, for example, the cofactors are ± determinants of the 3×3 matrices.

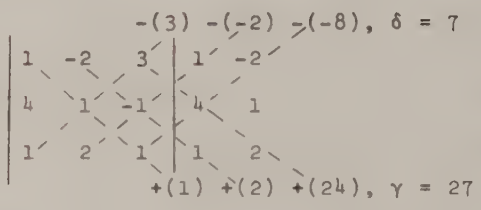
Another alternative of defining a det A, 3×3, is by using the rule of Sarrus. The figure below portrays the steps needed for defining det A.



$$\det A = (\gamma)+(\delta)$$

Example

By using the data of the above example define det A,



$$\det A = \gamma+\delta = 27+7 = 34$$

3.1 Application of determinants to simultaneous equations

We have seen in section 1 that the general expression for the system of a pair of simultaneous equations is,

$$a_1x+b_1y = c_1$$
$$a_2x+b_2y = c_2$$

A possible method of solving the system is by the application of determinants (Gramer's rule), i.e.,

$$x = \frac{\begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}} = \frac{c_1b_2-c_2b_1}{a_1b_2-a_2b_1}$$

$$y = \frac{\begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}} = \frac{a_1c_2-a_2c_1}{a_1b_2-a_2b_1}$$

Example

Solve the system (see section 1),

$$2x-5y = 19$$
$$3x+4y = -6$$

$$x = \frac{\begin{vmatrix} 19 & -5 \\ -6 & 4 \end{vmatrix}}{\begin{vmatrix} 2 & -5 \\ 3 & 4 \end{vmatrix}} = \frac{76-30}{8+15} = 2$$

$$y = \frac{\begin{vmatrix} 2 & 19 \\ 3 & -6 \end{vmatrix}}{\begin{vmatrix} 2 & -5 \\ 3 & 4 \end{vmatrix}} = \frac{-12-57}{23} = -3$$

Hence the solution is (2, -3)

Gramer's rule can be used to solve systems of simultaneous linear equations described above. In general, to solve the system,

$$\begin{array}{l|l}
(1) & a_1x_1+a_2x_2+ \ldots +a_nx_n = a_o, \quad \text{Not all } a_1, a_2 \ldots a_n = 0 \\
(2) & b_1x_1+b_2x_2+ \ldots +b_nx_n = b_o, \quad \text{Not all } b_1, b_2 \ldots b_n = 0 \\
(3) & \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad . \\
\ldots & \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad . \\
\ldots & \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad . \\
(n) & \lambda_1x_1+\lambda_2x_2+ \ldots +\lambda_nx_n = \lambda_o, \quad \text{Not all } \lambda_1, \lambda_2 \ldots \lambda_n = 0
\end{array}$$

We have,

$$x_1 = \begin{vmatrix} a_0 & a_2 & a_3 & \ldots & a_n \\ b_0 & b_2 & b_3 & \ldots & b_n \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ \lambda_0 & \lambda_2 & \lambda_3 & \ldots & \lambda_n \end{vmatrix} \div \begin{vmatrix} a_1 & a_2 & a_3 & \ldots & a_n \\ b_1 & b_2 & b_3 & \ldots & b_n \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ \lambda_1 & \lambda_2 & \lambda_3 & \ldots & \lambda_n \end{vmatrix}$$

$$x_2 = \begin{vmatrix} a_1 & a_0 & a_3 & \ldots & a_n \\ b_1 & b_0 & b_3 & \ldots & b_n \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ \lambda_1 & \lambda_0 & \lambda_3 & \ldots & \lambda_n \end{vmatrix} \div \begin{vmatrix} a_1 & a_2 & a_3 & \ldots & a_n \\ b_1 & b_2 & b_3 & \ldots & b_n \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \cdot & \ldots & \cdot \\ \lambda_1 & \lambda_2 & \lambda_3 & \ldots & \lambda_n \end{vmatrix}$$

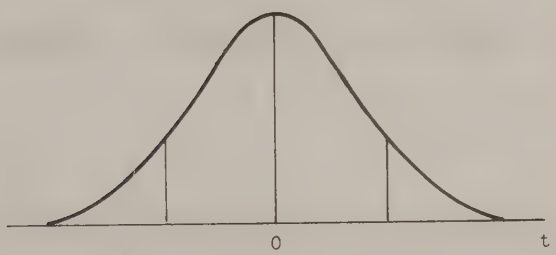
$$\ldots \quad \begin{vmatrix} \cdot & \cdot & \cdot & \ldots & \cdot \end{vmatrix} \quad \begin{vmatrix} \cdot & \cdot & \cdot & \ldots & \cdot \end{vmatrix}$$

$$\ldots \quad \begin{vmatrix} \cdot & \cdot & \cdot & \ldots & \cdot \end{vmatrix} \quad \begin{vmatrix} \cdot & \cdot & \cdot & \ldots & \cdot \end{vmatrix}$$

$$x_n = \begin{vmatrix} a_1 & a_2 & a_3 & \cdots & a_0 \\ b_1 & b_2 & b_3 & \cdots & b_0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \lambda_1 & \lambda_2 & \lambda_3 & \cdots & \lambda_0 \end{vmatrix} \div \begin{vmatrix} a_1 & a_2 & a_3 & \cdots & a_n \\ b_1 & b_2 & b_3 & \cdots & b_n \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \lambda_1 & \lambda_2 & \lambda_3 & \cdots & \lambda_n \end{vmatrix}$$

APPENDIX II - STATISTICAL TABLES

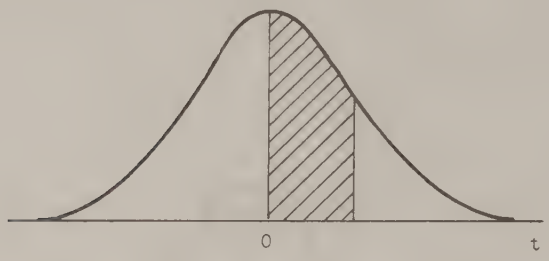
1. Ordinates of the standard normal curve[1/]



t	f(t)	t	f(t)
0.0	0.39894	2.1	0.04398
0.1	0.39695	2.2	0.03547
0.2	0.39104	2.3	0.02833
0.3	0.38129	2.4	0.02239
0.4	0.36827	2.5	0.01753
0.5	0.35207	2.6	0.01358
0.6	0.33322	2.7	0.01042
0.7	0.31225	2.8	0.00792
0.8	0.28969	2.9	0.00595
0.9	0.26609	3.0	0.00443
1.0	0.24197	3.1	0.00327
1.1	0.21785	3.2	0.00238
1.2	0.19419	3.3	0.00172
1.3	0.17137	3.4	0.00123
1.4	0.14973	3.5	0.00087
1.5	0.12952	3.6	0.00061
1.6	0.11092	3.7	0.00042
1.7	0.09405	3.8	0.00029
1.8	0.07895	3.9	0.00020
1.9	0.06562	4.0	0.00013
2.0	0.05399		

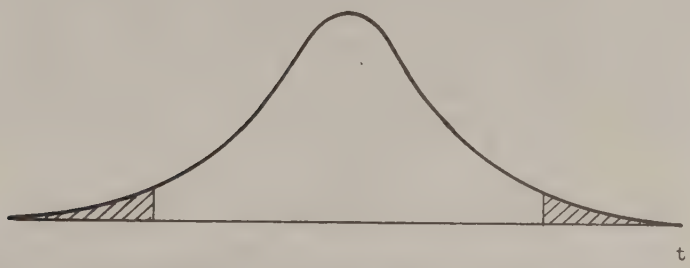
1/ Another tabulation of the values of ordinates is to multiply the values of the above table by $\sqrt{2\pi} = 2.5066$.

2. Areas under the normal curve from 0 to t (shown shaded is A(t))



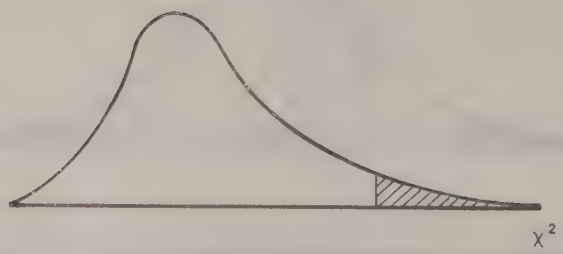
t	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0159	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3233
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3718	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4430	.4441
1.6	.4452	.4463	.4474	.4485	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4758	.4762	.4767
2.0	.4773	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4865	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4980	.4980	.4981
2.9	.4981	.4982	.4984	.4984	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4988	.4989	.4989	.4989	.4990

3. Table of Students t-distribution



Degrees of freedom	.01	.02	.05	.1	.2	.4	.6	.8	.9
1	63.657	31.821	12.706	6.314	3.078	1.376	.727	.325	.158
2	9.925	6.965	4.303	2.920	1.886	1.061	.617	.289	.142
3	5.841	4.541	3.182	2.353	1.638	.978	.584	.277	.137
4	4.604	3.747	2.776	2.132	1.533	.941	.569	.271	.134
5	4.032	3.365	2.571	2.015	1.476	.920	.559	.267	.132
6	3.707	3.143	2.447	1.943	1.440	.906	.553	.265	.131
7	3.499	2.998	2.365	1.895	1.415	.896	.549	.263	.130
8	3.355	2.896	2.306	1.860	1.397	.889	.546	.262	.130
9	3.250	2.821	2.262	1.833	1.383	.883	.543	.261	.129
10	3.169	2.764	2.228	1.812	1.372	.879	.542	.260	.129
11	3.106	2.718	2.201	1.796	1.363	.876	.540	.260	.129
12	3.055	2.681	2.179	1.782	1.356	.873	.539	.259	.128
13	3.012	2.650	2.160	1.771	1.350	.870	.538	.259	.128
14	2.977	2.624	2.145	1.761	1.345	.868	.537	.258	.128
15	2.947	2.602	2.131	1.753	1.341	.866	.536	.258	.128
16	2.921	2.583	2.120	1.746	1.337	.865	.535	.258	.128
17	2.898	2.567	2.110	1.740	1.333	.863	.534	.257	.128
18	2.878	2.552	2.101	1.734	1.330	.862	.534	.257	.127
19	2.861	2.539	2.093	1.729	1.328	.861	.533	.257	.127
20	2.845	2.528	2.086	1.725	1.325	.860	.533	.257	.127
21	2.831	2.518	2.080	1.721	1.323	.859	.532	.257	.127
22	2.819	2.508	2.074	1.717	1.321	.858	.532	.256	.127
23	2.807	2.500	2.069	1.714	1.319	.858	.532	.256	.127
24	2.797	2.492	2.064	1.711	1.318	.857	.531	.256	.127
25	2.787	2.485	2.060	1.708	1.316	.856	.531	.256	.127
26	2.779	2.479	2.056	1.706	1.315	.856	.531	.256	.127
27	2.771	2.473	2.052	1.703	1.314	.855	.531	.256	.127
28	2.763	2.467	2.048	1.701	1.313	.855	.530	.256	.127
29	2.756	2.462	2.045	1.699	1.311	.854	.530	.256	.127
30	2.750	2.457	2.042	1.697	1.310	.854	.530	.256	.127
∞	2.576	2.326	1.960	1.645	1.282	.842	.524	.253	.126

4. Table of χ^2 distribution



Degrees of freedom	P = 0.99	0.95	0.90	0.10	0.05	0.02	0.01
1	0.000157	0.00393	0.0158	2.706	3.841	5.412	6.635
2	0.0201	0.103	0.211	4.605	5.991	7.824	9.210
3	0.115	0.352	0.584	6.251	7.815	9.837	11.345
4	0.297	0.711	1.064	7.779	9.488	11.668	13.277
5	0.554	1.145	1.610	9.236	11.070	13.388	15.086
6	0.872	1.635	2.204	10.645	12.592	15.033	16.812
7	1.239	2.167	2.833	12.017	14.067	16.622	18.475
8	1.646	2.733	3.490	13.362	15.507	18.168	20.090
9	2.088	3.325	4.168	14.684	16.919	19.679	21.666
10	2.558	3.940	4.865	15.987	18.307	21.161	23.209
11	3.053	4.575	5.578	17.275	19.675	22.618	24.725
12	3.571	5.226	6.304	18.549	21.026	24.054	26.217
13	4.107	5.892	7.042	19.812	22.362	25.472	27.688
14	4.660	6.571	7.790	21.064	23.685	26.873	29.141
15	5.229	7.261	8.547	22.307	24.996	28.259	30.578
16	5.812	7.962	9.312	23.542	26.296	29.633	32.000
17	6.408	8.672	10.085	24.769	27.587	30.995	33.409
18	7.015	9.390	10.865	25.989	28.869	32.346	34.805
19	7.633	10.117	11.651	27.204	30.144	33.687	36.191
20	8.260	10.851	12.443	28.412	31.410	35.020	37.566
21	8.897	11.591	13.240	29.615	32.671	36.343	38.932
22	9.542	12.338	14.041	30.813	33.924	37.659	40.289
23	10.196	13.091	14.848	32.007	35.172	38.968	41.638
24	10.856	13.848	15.659	33.196	36.415	40.270	42.980
25	11.524	14.611	16.473	34.382	37.652	41.566	44.314
26	12.198	15.379	17.292	35.563	38.885	42.856	45.642
27	12.879	16.151	18.114	36.741	40.113	44.140	46.963
28	13.565	16.928	18.939	37.916	41.337	45.419	48.278
29	14.256	17.708	19.768	39.087	42.557	46.693	49.588
30	14.953	18.493	20.599	40.256	43.773	47.962	50.892

5. Table of z for values of r from .00 to 1.00, $z = \frac{1}{2}\log_e \frac{(1+r)}{(1-r)}$

r	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0100	.0200	.0300	.0400	.0500	.0601	.0701	.0802	.0902
.1	.1003	.1104	.1206	.1307	.1409	.1511	.1614	.1717	.1820	.1923
.2	.2027	.2132	.2237	.2342	.2448	.2554	.2661	.2769	.2877	.2986
.3	.3095	.3205	.3316	.3428	.3541	.3654	.3769	.3884	.4001	.4118
.4	.4236	.4356	.4477	.4599	.4722	.4847	.4973	.5101	.5230	.5361
.5	.5493	.5627	.5763	.5901	.6042	.6184	.6328	.6475	.6625	.6777
.6	.6931	.7089	.7250	.7414	.7582	.7753	.7928	.8107	.8291	.8480
.7	.8673	.8872	.9076	.9287	.9505	.9730	.9962	1.0203	1.0454	1.0714
.8	1.0986	1.1270	1.1568	1.1881	1.2212	1.2562	1.2933	1.3331	1.3758	1.4219
.9	1.4722	1.5275	1.5890	1.6584	1.7380	1.8318	1.9459	2.0923	2.2976	2.6467

6. Table of r for values of z from 0 to 5*

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0100	.0200	.0300	.0400	.0500	.0599	.0699	.0798	.0898
.1	.0997	.1096	.1194	.1293	.1391	.1489	.1587	.1684	.1781	.1878
.2	.1974	.2070	.2165	.2260	.2355	.2449	.2543	.2636	.2729	.2821
.3	.2913	.3004	.3095	.3185	.3275	.3364	.3452	.3540	.3627	.3714
.4	.3800	.3885	.3969	.4053	.4136	.4219	.4301	.4382	.4462	.4542
.5	.4621	.4700	.4777	.4854	.4930	.5005	.5080	.5154	.5227	.5299
.6	.5370	.5441	.5511	.5581	.5649	.5717	.5784	.5850	.5915	.5980
.7	.6044	.6107	.6169	.6231	.6291	.6352	.6411	.6469	.6527	.6584
.8	.6640	.6696	.6751	.6805	.6858	.6911	.6963	.7014	.7064	.7114
.9	.7163	.7211	.7259	.7306	.7352	.7398	.7443	.7487	.7531	.7574
1.0	.7616	.7658	.7699	.7739	.7779	.7818	.7857	.7895	.7932	.7969
1.1	.8005	.8041	.8076	.8110	.8144	.8178	.8210	.8243	.8275	.8306
1.2	.8337	.8367	.8397	.8426	.8455	.8483	.8511	.8538	.8565	.8591
1.3	.8617	.8643	.8668	.8693	.8717	.8741	.8764	.8787	.8810	.8832
1.4	.8854	.8875	.8896	.8917	.8937	.8957	.8977	.8996	.9015	.9033
1.5	.9052	.9069	.9087	.9104	.9121	.9138	.9154	.9170	.9186	.9202
1.6	.9217	.9232	.9246	.9261	.9275	.9289	.9302	.9316	.9329	.9342
1.7	.9354	.9367	.9379	.9391	.9402	.9414	.9425	.9436	.9447	.9458
1.8	.9468	.9478	.9488	.9498	.9508	.9518	.9527	.9536	.9545	.9554
1.9	.9562	.9571	.9579	.9587	.9595	.9603	.9611	.9619	.9626	.9633
2.0	.9640	.9647	.9654	.9661	.9668	.9674	.9680	.9687	.9693	.9699
2.1	.9705	.9710	.9716	.9722	.9727	.9732	.9738	.9743	.9748	.9753
2.2	.9757	.9762	.9767	.9771	.9776	.9780	.9785	.9789	.9793	.9797
2.3	.9801	.9805	.9809	.9812	.9816	.9820	.9823	.9827	.9830	.9834
2.4	.9837	.9840	.9843	.9846	.9849	.9852	.9855	.9858	.9861	.9863
2.5	.9866	.9869	.9871	.9874	.9876	.9879	.9881	.9884	.9886	.9888
2.6	.9890	.9892	.9895	.9897	.9899	.9901	.9903	.9905	.9906	.9908
2.7	.9910	.9912	.9914	.9915	.9917	.9919	.9920	.9922	.9923	.9925
2.8	.9926	.9928	.9929	.9931	.9932	.9933	.9935	.9936	.9937	.9938
2.9	.9940	.9941	.9942	.9943	.9944	.9945	.9946	.9947	.9949	.9950
3.0	.9951									
4.0	.9993									
5.0	.9999									

* The figures in the body of the table are values of r corresponding to z values read from the scales on the left and top of the table.

7. Table

5% (small type) and 1% (large type)

n_2	n_1 degrees of freedom (fo											
	1	2	3	4	5	6	7	8	9	10	11	12
1	161 4052	200 4999	216 5403	225 5625	230 5764	234 5859	237 5928	239 5981	241 6022	242 6056	243 6082	244 6106
2	18.51 98.49	19.00 99.00	19.16 99.17	19.25 99.25	19.30 99.30	19.33 99.33	19.36 99.34	19.37 99.36	19.38 99.38	19.39 99.40	19.40 99.41	19.41 99.42
3	10.13 34.12	9.55 30.82	9.28 29.46	9.12 28.71	9.01 28.24	8.94 27.91	8.88 27.67	8.84 27.49	8.81 27.34	8.78 27.23	8.76 27.13	8.74 27.05
4	7.71 21.20	6.94 18.00	6.59 16.69	6.39 15.98	6.26 15.52	6.16 15.21	6.09 14.98	6.04 14.80	6.00 14.66	5.96 14.54	5.93 14.45	5.91 14.37
5	6.61 16.26	5.79 13.27	5.41 12.06	5.19 11.39	5.05 10.97	4.95 10.67	4.88 10.45	4.82 10.27	4.78 10.15	4.74 10.05	4.70 9.96	4.68 9.89
6	5.99 13.74	5.14 10.92	4.76 9.78	4.53 9.15	4.39 8.75	4.28 8.47	4.21 8.26	4.15 8.10	4.10 7.98	4.06 7.87	4.03 7.79	4.00 7.72
7	5.59 12.25	4.74 9.55	4.35 8.45	4.12 7.85	3.97 7.46	3.87 7.19	3.79 7.00	3.73 6.84	3.68 6.71	3.63 6.62	3.60 6.54	3.57 6.47
8	5.32 11.26	4.46 8.65	4.07 7.59	3.84 7.01	3.69 6.63	3.58 6.37	3.50 6.19	3.44 6.03	3.39 5.91	3.34 5.82	3.31 5.74	3.28 5.67
9	5.12 10.56	4.26 8.02	3.86 6.99	3.63 6.42	3.48 6.06	3.37 5.80	3.29 5.62	3.23 5.47	3.18 5.35	3.13 5.26	3.10 5.18	3.07 5.11
10	4.96 10.04	4.10 7.56	3.71 6.55	3.48 5.99	3.33 5.64	3.22 5.39	3.14 5.21	3.07 5.06	3.02 4.95	2.97 4.85	2.94 4.78	2.91 4.71
11	4.84 9.65	3.98 7.20	3.59 6.22	3.36 5.67	3.20 5.32	3.09 5.07	3.01 4.88	2.95 4.74	2.90 4.63	2.86 4.54	2.82 4.46	2.79 4.40
12	4.75 9.33	3.88 6.93	3.49 5.95	3.26 5.41	3.11 5.06	3.00 4.82	2.92 4.65	2.85 4.50	2.80 4.39	2.76 4.30	2.72 4.22	2.69 4.16
13	4.67 9.07	3.80 6.70	3.41 5.74	3.18 5.20	3.02 4.86	2.92 4.62	2.84 4.44	2.77 4.30	2.72 4.19	2.67 4.10	2.63 4.02	2.60 3.96
14	4.60 8.86	3.74 6.51	3.34 5.56	3.11 5.03	2.96 4.69	2.85 4.46	2.77 4.28	2.70 4.14	2.65 4.03	2.60 3.94	2.56 3.86	2.53 3.80
15	4.54 8.68	3.68 6.36	3.29 5.42	3.06 4.89	2.90 4.56	2.79 4.32	2.70 4.14	2.64 4.00	2.59 3.89	2.55 3.80	2.51 3.73	2.48 3.67
16	4.49 8.53	3.63 6.23	3.24 5.29	3.01 4.77	2.85 4.44	2.74 4.20	2.66 4.03	2.59 3.89	2.54 3.78	2.49 3.69	2.45 3.61	2.42 3.55
17	4.45 8.40	3.59 6.11	3.20 5.18	2.96 4.67	2.81 4.34	2.70 4.10	2.62 3.93	2.55 3.79	2.50 3.68	2.45 3.59	2.41 3.52	2.38 3.45
18	4.41 8.28	3.55 6.01	3.16 5.09	2.93 4.58	2.77 4.25	2.66 4.01	2.58 3.85	2.51 3.71	2.46 3.60	2.41 3.51	2.37 3.44	2.34 3.37
19	4.38 8.18	3.52 5.93	3.13 5.01	2.90 4.50	2.74 4.17	2.63 3.94	2.55 3.77	2.48 3.63	2.43 3.52	2.38 3.43	2.34 3.36	2.31 3.30
20	4.35 8.10	3.49 5.85	3.10 4.94	2.87 4.43	2.71 4.10	2.60 3.87	2.52 3.71	2.45 3.56	2.40 3.45	2.35 3.37	2.31 3.30	2.28 3.23
21	4.32 8.02	3.47 5.78	3.07 4.87	2.84 4.37	2.68 4.04	2.57 3.81	2.49 3.65	2.42 3.51	2.37 3.40	2.32 3.31	2.28 3.24	2.25 3.17
22	4.30 7.94	3.44 5.72	3.05 4.82	2.82 4.31	2.66 3.99	2.55 3.76	2.47 3.59	2.40 3.45	2.35 3.35	2.30 3.26	2.26 3.18	2.23 3.12
23	4.28 7.88	3.42 5.66	3.03 4.76	2.80 4.26	2.64 3.94	2.53 3.71	2.45 3.54	2.38 3.41	2.32 3.30	2.28 3.21	2.24 3.14	2.20 3.07
24	4.26 7.82	3.40 5.61	3.01 4.72	2.78 4.22	2.62 3.90	2.51 3.67	2.43 3.50	2.36 3.36	2.30 3.25	2.26 3.17	2.22 3.09	2.18 3.03
25	4.24 7.77	3.38 5.57	2.99 4.68	2.76 4.18	2.60 3.86	2.49 3.63	2.41 3.46	2.34 3.32	2.28 3.21	2.24 3.13	2.20 3.05	2.16 2.99
26	4.22 7.72	3.37 5.53	2.98 4.64	2.74 4.14	2.59 3.82	2.47 3.59	2.39 3.42	2.32 3.29	2.27 3.17	2.22 3.09	2.18 3.02	2.15 2.96
27	4.21 7.68	3.35 5.49	2.96 4.60	2.73 4.11	2.57 3.79	2.46 3.56	2.37 3.39	2.30 3.26	2.25 3.14	2.20 3.06	2.16 2.98	2.13 2.93
28	4.20 7.64	3.34 5.45	2.95 4.57	2.71 4.07	2.56 3.76	2.44 3.53	2.36 3.36	2.29 3.23	2.24 3.11	2.19 3.03	2.15 2.95	2.12 2.90
29	4.18 7.60	3.33 5.42	2.93 4.54	2.70 4.04	2.54 3.73	2.43 3.50	2.35 3.33	2.28 3.20	2.22 3.08	2.18 3.00	2.14 2.92	2.10 2.87
30	4.17 7.56	3.32 5.39	2.92 4.51	2.69 4.02	2.53 3.70	2.42 3.47	2.34 3.30	2.27 3.17	2.21 3.06	2.16 2.98	2.12 2.90	2.09 2.84
32	4.15 7.50	3.30 5.34	2.90 4.46	2.67 3.97	2.51 3.66	2.40 3.42	2.32 3.25	2.25 3.12	2.19 3.01	2.14 2.94	2.10 2.86	2.07 2.80
34	4.13 7.44	3.28 5.29	2.88 4.42	2.65 3.93	2.49 3.61	2.38 3.38	2.30 3.21	2.23 3.08	2.17 2.97	2.12 2.89	2.08 2.82	2.05 2.76
36	4.11 7.39	3.26 5.25	2.86 4.38	2.63 3.89	2.48 3.58	2.36 3.35	2.28 3.18	2.21 3.04	2.15 2.94	2.10 2.86	2.06 2.78	2.03 2.72
38	4.10 7.35	3.25 5.21	2.85 4.34	2.62 3.86	2.46 3.54	2.35 3.32	2.26 3.15	2.19 3.02	2.14 2.91	2.09 2.82	2.05 2.75	2.02 2.69
40	4.08 7.31	3.23 5.18	2.84 4.31	2.61 3.83	2.45 3.51	2.34 3.29	2.25 3.12	2.18 2.99	2.12 2.88	2.07 2.80	2.04 2.73	2.00 2.66
42	4.07 7.27	3.22 5.15	2.83 4.29	2.59 3.80	2.44 3.49	2.32 3.26	2.24 3.10	2.17 2.96	2.11 2.86	2.06 2.77	2.02 2.70	1.99 2.64
44	4.06 7.24	3.21 5.12	2.82 4.26	2.58 3.78	2.43 3.46	2.31 3.24	2.23 3.07	2.16 2.94	2.10 2.84	2.05 2.75	2.01 2.68	1.98 2.62
46	4.05 7.21	3.20 5.10	2.81 4.24	2.57 3.76	2.42 3.44	2.30 3.22	2.22 3.05	2.14 2.92	2.09 2.82	2.04 2.73	2.00 2.66	1.97 2.60
48	4.04 7.19	3.19 5.08	2.80 4.22	2.56 3.74	2.41 3.42	2.30 3.20	2.21 3.04	2.14 2.90	2.08 2.80	2.03 2.71	1.99 2.64	1.96 2.58
50	4.03 7.17	3.18 5.06	2.79 4.20	2.56 3.72	2.40 3.41	2.29 3.18	2.20 3.02	2.13 2.88	2.07 2.78	2.02 2.70	1.98 2.62	1.95 2.56
55	4.02 7.12	3.17 5.01	2.78 4.16	2.54 3.68	2.38 3.37	2.27 3.15	2.18 2.98	2.11 2.85	2.05 2.75	2.00 2.66	1.97 2.59	1.93 2.53
60	4.00 7.08	3.15 4.98	2.76 4.13	2.52 3.65	2.37 3.34	2.25 3.12	2.17 2.95	2.10 2.82	2.04 2.72	1.99 2.63	1.95 2.56	1.92 2.50
65	3.99 7.04	3.14 4.95	2.75 4.10	2.51 3.62	2.36 3.31	2.24 3.09	2.15 2.93	2.08 2.79	2.02 2.70	1.98 2.61	1.94 2.54	1.90 2.47
70	3.98 7.01	3.13 4.92	2.74 4.08	2.50 3.60	2.35 3.29	2.23 3.07	2.14 2.91	2.07 2.77	2.01 2.67	1.97 2.59	1.93 2.51	1.89 2.45
80	3.96 6.96	3.11 4.88	2.72 4.04	2.48 3.56	2.33 3.25	2.21 3.04	2.12 2.87	2.05 2.74	1.99 2.64	1.95 2.55	1.91 2.48	1.88 2.41
100	3.94 6.90	3.09 4.82	2.70 3.98	2.46 3.51	2.30 3.20	2.19 2.99	2.10 2.82	2.03 2.69	1.97 2.59	1.92 2.51	1.88 2.43	1.85 2.36
125	3.92 6.84	3.07 4.78	2.68 3.94	2.44 3.47	2.29 3.17	2.17 2.95	2.08 2.79	2.01 2.65	1.95 2.56	1.90 2.47	1.86 2.40	1.83 2.33
150	3.91 6.81	3.06 4.75	2.67 3.91	2.43 3.44	2.27 3.14	2.16 2.92	2.07 2.76	2.00 2.62	1.94 2.53	1.89 2.44	1.85 2.37	1.82 2.30
200	3.89 6.76	3.04 4.71	2.65 3.88	2.41 3.41	2.26 3.11	2.14 2.90	2.05 2.73	1.98 2.60	1.92 2.50	1.87 2.41	1.83 2.34	1.80 2.28
400	3.86 6.70	3.02 4.66	2.62 3.83	2.39 3.36	2.23 3.06	2.12 2.85	2.03 2.69	1.96 2.55	1.90 2.46	1.85 2.37	1.81 2.29	1.78 2.23
1000	3.85 6.66	3.00 4.62	2.61 3.80	2.38 3.34	2.22 3.04	2.10 2.82	2.02 2.66	1.95 2.53	1.89 2.43	1.84 2.34	1.80 2.26	1.76 2.20
∞	3.84 6.64	2.99 4.60	2.60 3.78	2.37 3.32	2.21 3.02	2.09 2.80	2.01 2.64	1.94 2.51	1.88 2.41	1.83 2.32	1.79 2.24	1.75 2.18

of F

points for the distribution of F

greater mean square)												n_2
14	16	20	24	30	40	50	75	100	200	500	∞	
245 6142	246 6169	248 6208	249 6234	250 6258	251 6282	252 6302	253 6323	253 6334	254 6352	254 6361	254 6366	1
19.42 99.43	19.43 99.44	19.44 99.45	19.45 99.46	19.46 99.47	19.47 99.48	19.47 99.48	19.48 99.49	19.49 99.49	19.49 99.49	19.50 99.50	19.50 99.50	2
8.71 26.92	8.69 26.83	8.66 26.69	8.64 26.60	8.62 26.50	8.60 26.41	8.58 26.35	8.57 26.27	8.56 26.23	8.54 26.18	8.54 26.14	8.53 26.12	3
5.87 14.24	5.84 14.15	5.80 14.02	5.77 13.93	5.74 13.83	5.71 13.74	5.70 13.69	5.68 13.61	5.66 13.57	5.65 13.52	5.64 13.48	5.63 13.46	4
4.64 9.77	4.60 9.68	4.56 9.55	4.53 9.47	4.50 9.38	4.46 9.29	4.44 9.24	4.42 9.17	4.40 9.13	4.38 9.07	4.37 9.04	4.36 9.02	5
3.96 7.60	3.92 7.52	3.87 7.39	3.84 7.31	4.81 7.23	3.77 7.14	3.75 7.09	3.72 7.02	3.71 6.99	3.69 6.94	3.68 6.90	3.67 6.88	6
3.52 6.35	3.49 6.27	3.44 6.15	3.41 6.07	3.38 5.98	3.34 5.90	3.32 5.85	3.29 5.78	3.28 5.75	3.25 5.70	3.24 5.67	3.23 5.65	7
3.23 5.56	3.20 5.48	3.15 5.36	3.12 5.28	3.08 5.20	3.05 5.11	3.03 5.06	3.00 5.00	2.98 4.96	2.96 4.91	2.94 4.88	2.93 4.86	8
3.02 5.00	2.98 4.92	2.93 4.80	2.90 4.73	2.86 4.64	2.82 4.56	2.80 4.51	2.77 4.45	2.76 4.41	2.73 4.36	2.72 4.33	2.71 4.31	9
2.86 4.60	2.82 4.52	2.77 4.41	2.74 4.33	2.70 4.25	2.67 4.17	2.64 4.12	2.61 4.05	2.59 4.01	2.56 3.96	2.55 3.93	2.54 3.91	10
2.74 4.29	2.70 4.21	2.65 4.10	2.61 4.02	2.57 3.94	2.53 3.86	2.50 3.80	2.47 3.74	2.45 3.70	2.42 3.66	2.41 3.62	2.40 3.60	11
2.64 4.05	2.60 3.98	2.54 3.86	2.50 3.78	2.46 3.70	2.42 3.61	2.40 3.56	2.36 3.49	2.35 3.46	2.32 3.41	2.31 3.38	2.30 3.36	12
2.55 3.85	2.51 3.78	2.46 3.67	2.42 3.59	2.38 3.51	2.34 3.42	2.32 3.37	2.28 3.30	2.26 3.27	2.24 3.21	2.22 3.18	2.21 3.16	13
2.48 3.70	2.44 3.62	2.39 3.51	2.35 3.43	2.31 3.34	2.27 3.26	2.24 3.21	2.21 3.14	2.19 3.11	2.16 3.06	2.14 3.02	2.13 3.00	14
2.43 3.56	2.39 3.48	2.33 3.36	2.29 3.29	2.25 3.20	2.21 3.12	2.18 3.07	2.15 3.00	2.12 2.97	2.10 2.92	2.08 2.89	2.07 2.87	15
2.37 3.45	2.33 3.37	2.28 3.25	2.24 3.18	2.20 3.10	2.16 3.01	2.13 2.96	2.09 2.89	2.07 2.86	2.04 2.80	2.02 2.77	2.01 2.75	16
2.33 3.35	2.29 3.27	2.23 3.16	2.19 3.08	2.15 3.00	2.11 2.92	2.08 2.86	2.04 2.79	2.02 2.76	1.99 2.70	1.97 2.67	1.96 2.65	17
2.29 3.27	2.25 3.19	2.19 3.07	2.15 3.00	2.11 2.91	2.07 2.83	2.04 2.78	2.00 2.71	1.98 2.68	1.95 2.62	1.93 2.59	1.92 2.57	18
2.26 3.19	2.21 3.12	2.15 3.00	2.11 2.92	2.07 2.84	2.02 2.76	2.00 2.70	1.96 2.63	1.94 2.60	1.91 2.54	1.90 2.51	1.88 2.49	19
2.23 3.13	2.18 3.05	2.12 2.94	2.08 2.86	2.04 2.77	1.99 2.69	1.96 2.63	1.92 2.56	1.90 2.53	1.87 2.47	1.85 2.44	1.84 2.42	20
2.20 3.07	2.15 2.99	2.09 2.88	2.05 2.80	2.00 2.72	1.96 2.63	1.93 2.58	1.89 2.51	1.87 2.47	1.84 2.42	1.82 2.38	1.81 2.36	21
2.18 3.02	2.13 2.94	2.07 2.83	2.03 2.75	1.98 2.67	1.93 2.58	1.91 2.53	1.87 2.46	1.84 2.42	1.81 2.37	1.80 2.33	1.78 2.31	22
2.14 2.97	2.10 2.89	2.04 2.78	2.00 2.70	1.96 2.62	1.91 2.53	1.88 2.48	1.84 2.41	1.82 2.37	1.79 2.32	1.77 2.28	1.76 2.26	23
2.13 2.93	2.09 2.85	2.02 2.74	1.98 2.66	1.94 2.58	1.89 2.49	1.86 2.44	1.82 2.36	1.80 2.33	1.76 2.27	1.74 2.23	1.73 2.21	24
2.11 2.89	2.06 2.81	2.00 2.70	1.96 2.62	1.92 2.54	1.87 2.45	1.84 2.40	1.80 2.32	1.77 2.29	1.74 2.23	1.72 2.19	1.71 2.17	25
2.10 2.86	2.05 2.77	1.99 2.66	1.95 2.58	1.90 2.50	1.85 2.41	1.82 2.36	1.78 2.28	1.76 2.25	1.72 2.19	1.70 2.15	1.69 2.13	26
2.08 2.83	2.03 2.74	1.97 2.63	1.93 2.55	1.88 2.47	1.84 2.38	1.80 2.33	1.76 2.25	1.74 2.21	1.71 2.16	1.68 2.12	1.67 2.10	27
2.06 2.80	2.02 2.71	1.96 2.60	1.91 2.52	1.87 2.44	1.81 2.35	1.78 2.30	1.75 2.22	1.72 2.18	1.69 2.13	1.67 2.09	1.65 2.06	28
2.05 2.77	2.00 2.68	1.94 2.57	1.90 2.49	1.85 2.41	1.80 2.32	1.77 2.27	1.73 2.19	1.71 2.15	1.68 2.10	1.65 2.06	1.64 2.03	29
2.04 2.74	1.99 2.66	1.93 2.55	1.89 2.47	1.84 2.38	1.79 2.29	1.76 2.24	1.72 2.16	1.69 2.13	1.66 2.07	1.64 2.03	1.62 2.01	30
2.02 2.70	1.97 2.62	1.91 2.51	1.86 2.42	1.82 2.34	1.76 2.25	1.74 2.20	1.69 2.12	1.67 2.08	1.64 2.02	1.61 1.98	1.59 1.96	32
2.00 2.66	1.95 2.58	1.89 2.47	1.84 2.38	1.80 2.30	1.74 2.21	1.71 2.15	1.67 2.08	1.64 2.04	1.61 1.98	1.59 1.94	1.57 1.91	34
1.98 2.62	1.93 2.54	1.87 2.43	1.82 2.35	1.78 2.26	1.72 2.17	1.69 2.12	1.65 2.04	1.62 2.00	1.59 1.94	1.56 1.90	1.55 1.87	36
1.96 2.59	1.92 2.51	1.85 2.40	1.80 2.32	1.76 2.22	1.71 2.14	1.67 2.08	1.63 2.00	1.60 1.97	1.57 1.90	1.54 1.86	1.53 1.84	38
1.95 2.56	1.90 2.49	1.84 2.37	1.79 2.29	1.74 2.20	1.69 2.11	1.66 2.05	1.61 1.97	1.59 1.94	1.55 1.88	1.53 1.84	1.51 1.81	40
1.94 2.54	1.89 2.46	1.82 2.35	1.78 2.26	1.73 2.17	1.68 2.08	1.64 2.02	1.60 1.94	1.57 1.91	1.54 1.85	1.51 1.80	1.49 1.78	42
1.92 2.52	1.88 2.44	1.81 2.32	1.76 2.24	1.72 2.15	1.66 2.06	1.63 2.00	1.58 1.92	1.56 1.88	1.52 1.82	1.50 1.78	1.48 1.75	44
1.91 2.50	1.87 2.42	1.80 2.30	1.75 2.22	1.71 2.13	1.65 2.04	1.62 1.98	1.57 1.90	1.54 1.86	1.51 1.80	1.48 1.76	1.46 1.72	46
1.90 2.48	1.86 2.40	1.79 2.28	1.74 2.20	1.70 2.11	1.64 2.02	1.61 1.96	1.56 1.88	1.53 1.84	1.50 1.78	1.47 1.73	1.45 1.70	48
1.90 2.46	1.85 2.39	1.78 2.27	1.74 2.19	1.69 2.10	1.63 2.00	1.60 1.94	1.55 1.86	1.52 1.82	1.48 1.76	1.46 1.71	1.44 1.68	50
1.88 2.43	1.83 2.35	1.76 2.23	1.72 2.15	1.67 2.06	1.61 1.98	1.58 1.90	1.52 1.82	1.50 1.78	1.46 1.71	1.43 1.66	1.41 1.64	55
1.86 2.40	1.81 2.32	1.75 2.20	1.70 2.12	1.65 2.03	1.59 1.93	1.56 1.87	1.50 1.79	1.48 1.74	1.44 1.68	1.41 1.63	1.39 1.60	60
1.85 2.37	1.80 2.30	1.73 2.18	1.68 2.09	1.63 2.00	1.57 1.90	1.54 1.84	1.49 1.76	1.46 1.71	1.42 1.64	1.39 1.60	1.37 1.56	65
1.84 2.35	1.79 2.28	1.72 2.15	1.67 2.07	1.62 1.98	1.56 1.88	1.53 1.82	1.47 1.74	1.45 1.69	1.40 1.62	1.37 1.56	1.35 1.53	70
1.82 2.32	1.77 2.24	1.71 2.11	1.65 2.03	1.60 1.94	1.54 1.84	1.51 1.78	1.45 1.70	1.42 1.65	1.38 1.57	1.35 1.52	1.32 1.49	80
1.79 2.26	1.75 2.19	1.68 2.06	1.63 1.98	1.57 1.89	1.51 1.79	1.48 1.73	1.42 1.64	1.39 1.59	1.34 1.51	1.30 1.46	1.28 1.43	100
1.77 2.23	1.73 2.15	1.65 2.03	1.60 1.94	1.55 1.85	1.49 1.75	1.46 1.68	1.39 1.59	1.36 1.54	1.31 1.46	1.27 1.40	1.25 1.37	125
1.76 2.20	1.71 2.12	1.64 2.00	1.59 1.91	1.54 1.83	1.47 1.72	1.44 1.66	1.37 1.56	1.34 1.51	1.29 1.43	1.25 1.37	1.22 1.33	150
1.74 2.17	1.69 2.09	1.62 1.97	1.57 1.88	1.52 1.79	1.45 1.69	1.42 1.62	1.35 1.53	1.32 1.48	1.26 1.39	1.22 1.33	1.19 1.28	200
1.72 2.12	1.67 2.04	1.60 1.92	1.54 1.84	1.49 1.74	1.42 1.64	1.38 1.57	1.32 1.47	1.28 1.42	1.22 1.32	1.16 1.24	1.13 1.19	400
1.70 2.09	1.65 2.01	1.58 1.89	1.53 1.81	1.47 1.71	1.41 1.61	1.36 1.54	1.30 1.44	1.26 1.38	1.19 1.28	1.13 1.19	1.08 1.11	1000

8. Table of random numbers

1	2	3	4	5	6	7	8	9	10	11	12
13	70	43	69	38	81	87	42	12	20	41	15
26	99	82	78	99	05	22	99	52	32	80	91
72	53	95	81	07	98	14	74	52	58	73	10
22	08	08	68	37	16	36	62	20	02	35	98
21	61	90	53	85	72	86	94	87	18	50	11
47	38	55	66	50	96	96	78	34	45	52	78
96	68	13	07	31	29	70	09	16	66	81	09
45	92	93	44	87	72	26	75	82	31	72	69
73	85	71	45	32	16	57	91	52	05	93	20
51	99	50	88	62	54	90	51	01	39	18	70
67	62	30	02	83	17	37	25	42	86	00	32
03	03	89	77	12	41	15	25	52	30	93	11
45	10	04	66	94	70	33	74	97	23	40	97
62	48	46	97	04	36	31	27	29	34	85	35
59	59	33	63	53	43	60	30	15	81	67	59
72	63	67	17	24	55	68	32	24	80	13	92
46	28	15	70	28	98	53	36	03	89	83	74
21	03	09	16	31	48	05	10	98	62	14	15
64	82	53	39	92	14	07	34	04	01	66	17
75	66	40	90	39	95	46	10	94	68	39	10
42	77	29	80	73	38	92	11	81	72	50	88
63	55	09	84	66	56	92	13	97	14	87	27
54	29	70	14	85	95	79	72	77	48	57	92
42	97	50	61	19	55	38	55	85	57	85	08
52	30	47	73	26	54	18	05	75	92	95	08
88	44	33	02	47	97	47	04	12	38	93	25
49	91	93	73	14	15	01	47	02	70	30	96
45	42	46	06	93	60	41	09	31	29	52	49
50	69	74	10	51	89	66	51	57	21	54	95
10	56	73	16	02	87	41	05	13	87	13	61

8. Table of random numbers (continued)

13	14	15	16	17	18	19	20	21	22	23	24
76	96	85	27	81	21	75	39	43	77	80	81
38	51	09	17	41	85	13	20	66	59	22	20
40	91	90	51	74	23	54	88	84	12	16	77
44	53	23	87	91	53	86	97	42	80	83	37
31	25	22	30	16	17	32	34	00	07	25	52
26	35	20	92	31	12	15	28	42	98	67	52
36	12	17	03	83	93	48	64	50	32	57	94
25	51	40	74	85	16	86	09	22	62	06	38
72	30	33	97	36	58	90	91	23	91	19	04
17	20	75	03	85	53	06	41	29	78	51	15
75	57	37	77	67	60	70	44	56	91	03	49
12	47	35	37	15	17	96	24	95	03	39	55
73	67	55	64	16	38	58	74	29	71	49	62
16	02	29	14	16	73	44	49	34	05	46	96
46	9	13	29	19	71	98	71	19	51	86	82
73	65	42	09	39	92	56	60	36	54	55	46
28	96	06	41	55	75	03	62	55	19	15	75
57	26	11	28	93	16	85	39	67	49	02	30
47	76	60	92	22	79	70	66	78	13	97	42
31	80	30	86	08	54	39	85	38	46	74	21
91	55	46	36	26	40	17	70	39	94	05	76
63	70	10	91	20	64	12	33	15	59	43	28
28	35	53	14	30	57	07	34	09	56	26	81
96	91	62	94	03	96	96	17	02	10	89	71
24	36	86	52	67	59	63	22	28	76	43	45
43	73	70	73	19	41	04	60	25	42	09	50
52	69	34	01	65	33	19	62	22	41	29	65
01	15	92	69	53	78	68	58	74	08	05	11
94	46	53	72	49	19	98	09	56	83	25	40
44	42	06	32	95	17	32	67	80	84	09	69
21	51	85	33	16	11	87	12	17	39	12	11
60	25	84	42	22	94	38	96	52	03	33	97
53	12	75	59	76	42	73	48	95	57	51	31
02	60	01	17	09	00	38	12	31	52	22	24
09	60	53	92	82	11	96	03	47	31	35	59

REFERENCES

Aitken, A.C., Statistical mathematics, 8th ed., Oliver and Boyd, Edinburgh 1957

Beyer, W.H., Handbook of tables for probability and statistics, 2nd ed., Chemical 1968 Rubber Co., Cleveland, Ohio

Croxton, F.E. et al., Applied general statistics, Sir Isaac Pitman and Sons Ltd., 1968 London

Ezekiel, M. and K.A. Fox, Methods of correlation and regression analysis: linear and 1966 curvilinear, 3rd ed., John Wiley and Sons, New York

Feller, W., An introduction to probability theory and its applications, John Wiley 1967 and Sons, New York

Fisher, R.A., The design of experiments, Oliver and Boyd, Edinburgh 1936

Goulden, C.H., Methods of statistical analysis, 2nd ed., John Wiley and Sons, New York 1952

Hoel, P.G., Introduction to mathematical statistics, John Wiley and Sons, New York 1954

Mood, A.M., Introduction to the theory of statistics, McGraw-Hill Book Co. Inc., 1950 New York

Mosteller, F. et al., Probability with statistical applications, Addison-Wesley, 1961 Reading, Mass.

Quenouille, M.H., Introductory statistics, Butterworth-Springer Ltd., London 1950

Rider, P.R., An introduction to modern statistical methods, John Wiley and Sons, New 1939 York

Schmid, C.F., Handbook of graphic presentation, Ronald Press Co., New York 1954

Snedecor, G.W. and W.G. Cochran, Statistical methods, The Iowa State Univ.Press, 1973 Ames, Iowa

Weatherburn, C.E., A first course in mathematical statistics, Univ.Press, Cambridge 1949

Yule, G.U. and M.G. Kendall, An introduction to the theory of statistics, Charles 1949 Griffin and Co. Ltd., London

LIST OF TECHNICAL REPORTS

The technical reports produced by the same author up to the present time, on the application of sampling techniques in fishery statistics (inland waters), are given under sections A-D in the list below. Copies can be obtained from Fishery Statistics Unit, Department of Fisheries, FAO/UN, Rome, Italy.

A: Statistical Studies (St.S.)

1. Sampling Techniques in Inland Fisheries with Special Reference to Volta Lake. St.S./1, FIO/SF/GHA/10, May 1970
2. Frame Surveys at Volta Lake. St.S./2, FIO/SF/GHA/10, March 1970
3. Yield Indices in Inland Fisheries with Special Reference to Volta Lake. St.S./3, FIO/SF/GHA/10, September 1971
4. Frame Survey at Kainji Lake. St.S./1, FI/SF/NIR/24, January 1971
5. The Yield Pattern at Lake Nasser. St.S./1, UNDP/SF/EGY/66/558, September 1972
6. The Yield Pattern at Kainji Lake. St.S./2, UNDP/SF/NIR/24, October 1972
7. Aerial Survey on the Lakes Malombe and Malawi. Analysis of the Results of the Survey. St.S./1, UNDP/SF/MLW/16, November 1972

†8. The Yield Pattern at Volta Lake. St.S./4, UNDP/SF/GHA/10, May 1973

**9. The Improvement of the Fisheries Statistical System at Lake Kossou. St.S./1, UNDP/SF/IVC/71/526, April 1973

**10. Coverage Check Survey of the Aerial Survey at Lake Kossou (CCS-AS). St.S./2, UNDP/SF/IVC/71/526, April 1973

11. The Improvement of the Fisheries Statistical System at Lake Tanganyika (Tanzania). St.S./1, UNDP/SF/URT/71/012, May 1973
12. Statistical Analysis of the Results of the Aerial Survey. St.S./1, Lake Victoria Fisheries Research Project, UNDP/SF/RAF/71/242, August 1973
13. Recent Trends in the Yield Pattern of Kainji Lake. St.S./3, UNDP/SF/NIR/24, April 1974

†14. Analysis of the Results of Frame Surveys 2 and 3 at Kainji Lake. St.S./4, UNDP/SF/NIR/24, April 1974

B: Statistical Efficiency (St.E.)

1. Efficiency of Different Sampling Methods for Large Scale Biological and Fishery Statistical Sample Surveys at Large African Lakes:
   1. Cove-Rotenone Sample Survey at Kariba Lake, St.E./1, UNDP/SF/ZAM/11, April 1972
2. Deck Sampling: An Assessment of a Pilot Trawling Survey on Lake Malawi (Malawi), St.E./2, UNDP/SF/MLW/16, February 1973

C: Fishery Statistical Surveys/Training (FSS.T.)

**1. Training Courses on Fishery Statistical Surveys (Inland Waters). FSS.T./1, UNDP/SF/ZAM/11, March 1973

** Also available in French.

† Will be available shortly.

D: Technical Reports

**1. Variation in Catchability of Fish with AVB and BOZO Gillnets (Lake Kossou), Report No.3, UNDP/SF/IVC/71/526, August 1973

**2. A Qualitative Evaluation of the Present Statistical System (SS) (Lake Kossou), Report No.8, UNDP/SF/IVC/71/526, August 1973

3. The Sampling Design of the Catch Assessment Survey (CAS), Lake Tanganyika (Tanzania), Report No.1, UNDP/SF/URT/71/012, December 1973

**4. Variations in Catchability of Fish with AVB and BOZO Gillnets (Lake Kossou): Analysis of the Results (Dry Season and in Coastal Areas). Report No.18, UNDP/SF/IVC/71/526, April 1974

E: Working Papers

*1. An Assessment of the Present Status of the Fishery at Lake Tanganyika (Burundi) (Traditional and artisanal fishery). Working Paper No.1, UNDP/SF/BDI/70/508, September 1973

*2. Yield Indices at Lake Tanganyika (Burundi). Working Paper No.2, UNDP/SF/BDI/70/508, September 1973

*3. Biological Sample Survey at Lake Tanganyika (Burundi): Statistical analysis of purse-seine fishery catches. Working Paper No.3, UNDP/SF/BDI/70/508, September 1973

** Also available in French.

* Will be available in French.

VARYSTAMPA
VIA MONTE DEL GALLO, 26/F
TEL. 630.674

MI/F3910/E/10.74/1/1000